MULTIVARIATE AND REGRESSION ANALYSIS BASED ON THE GEOMETRY OF DATA CLOUDS

Biman Chakraborty

Division of Theoretical Statistics & Mathematics

Indian Statistical Institute

203, B. T. Road, Calcutta 700035, INDIA

Thesis submitted to Indian Statistical Institute
in partial fulfilment of the requirements
for the award of the degree of
Doctor of Philosophy
CALCUTTA
1998

Acknowledgements

It appears to be customary to begin by thanking one's supervisor. But I find it superfluous, particularly since Professor Probal Chaudhuri means much more to me than merely my supervisor. After all, one normally does not thank his parents for holding his hands when he is taking his first faltering steps, for teaching him to walk. The materials of Chapters 2, 3 and parts of Chapter 4 have so far appeared as joint papers with Prof. Chaudhuri, which he has permitted me to include in this thesis.

I joined this Institute as an undergraduate student. The list of teachers who have taught me, nurtured me, encouraged me, made me what I am today, is naturally bound to read like the whole faculty list. To each of them I owe a lot. The same is also true for numerous friends and colleagues in the department or outside. Each of them gave me much needed moral support. I thank them all.

In the days of my research, I had the opportunity to communicate and discuss with Professor Hannu Oja, Professor Roger Koenker, Professor V. Koltchinski and Professor Jayanta K. Ghosh whose critical comments and suggestions on my research papers helped me a lot. My sincere thanks are to all of them. Parts of Chapters 2 and 3 appeared as a joint work with Prof. Hannu Oja and he too has kindly permitted me to include that in my thesis. I thank Dr. P. Bharati for making available some of the data-sets used in the thesis.

Finally, the nonacademic staff members of the Stat-Math Unit made it a lot easier for me by providing help whenever I needed. I am in their debt.

February, 1998 Biman Chakraborty

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Median is a natural estimate of location of a data set, and there are several versions of multivariate median studied in the literature, each of which is an interesting descriptive statistic for multivariate data and provides some nice geometric insights into the data cloud. One would expect that multidimensional median will be a natural estimate for the center of symmetry of a multivariate distribution. However, there is no unique concept of symmetry in multivariate problems. The center of symmetry can be defined in several ways there. For example, the d-dimensional random variable $\boldsymbol{X}$ is *spherically symmetric* about $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\boldsymbol{X} - \boldsymbol{\theta}$ and $\mathbf{O}(\boldsymbol{X} - \boldsymbol{\theta})$ are identically distributed for any orthogonal $d \times d$ matrix $\mathbf{O}$, and the distribution of a random variable $\boldsymbol{X}$ is said to be *elliptically symmetric* if there exists some positive definite matrix Σ such that $\Sigma^{-1/2}\boldsymbol{X}$ has a spherically symmetric distribution. One can relax the criteria of symmetry in order to define *central symmetry* as $\boldsymbol{X} - \boldsymbol{\theta}$ and $\boldsymbol{\theta} - \boldsymbol{X}$ having the same distribution. The concept of *angular symmetry* was suggested by Liu (1988). The random vector $\boldsymbol{X}$ is said to be angularly symmetric about $\boldsymbol{\theta}$ if the direction vector $(\boldsymbol{X} - \boldsymbol{\theta})/\|\boldsymbol{X} - \boldsymbol{\theta}\|$ is centrally symmetric about the origin.

From the definitions above, it is clear that all the notions of symmetry are sufficiently intuitive and worth studying. Any point $\boldsymbol{\theta}$ of spherical symmetry is a point of elliptical symmetry, and every point of elliptical symmetry is a point of central symmetry. In turn, any point of central symmetry is a point of angular symmetry. Closely related to the concept of a point of symmetry is the idea of the equivariance (or invariance) of a location estimate. For example, the univariate median is equivariant under monotone transformations of the real line, i.e. if $X_1, \ldots, X_n$ is a sample with median $\hat{\mu}(X_1, \ldots, X_n)$ and $h : \mathbb{R} \to \mathbb{R}$ is a monotone transformation, then

$$\hat{\mu}(h(X_1), \ldots, h(X_n)) = h(\hat{\mu}(X_1, \ldots, X_n)).$$

In the multivariate set up, one would expect an estimator of the point of spherical symmetry to be equivariant under the group of orthogonal transformations and translations. Similarly, an estimator of the point of elliptic symmetry should be equivariant under affine transformations of the data cloud. In subsequent sections, we will again discuss this property of equivariance while discussing some of the proposed multivariate medians in the literature.

Closely related to the concept of multivariate median is the concept of multivariate quantiles. Barnett (1976) has discussed in detail several methods for ordering multivariate data. Eddy (1983, 1985) approached the problem of multivariate quantiles through nested sequence of sets. Recently, Chaudhuri (1996) defined the concept of geometric quantiles, which generalizes the concept of spatial median to the quantile problem. According to Small (1990), an approach to quantiles can be based upon the fact that the maximization of the function $-E_F|\boldsymbol{X}-\mu|$ can be done by gradients, and which in a univariate situation reduces to the simple derivative. In higher dimensions, the gradient vector will typically point inwards to the center of the distribution with a length that is proportional to how exterior the location μ is (with respect to the distribution or its empirical analog) from the data set.

One of the early references to the concept of bivariate median can be found in Hayford (1902). He made a clear distinction between centroid of a spatial distribution (i.e. multivariate mean) and a median-like estimate of the center of a distribution. He suggested the vector of medians of orthogonal coordinates but clearly recognized that this higher dimensional analog of median is dependent on the choice of the orthogonal coordinate system. In other words, that median vector is not equivariant under orthogonal transformations or rotations. Gini and Galvani (1929) introduced the definition of spatial median. However, their work did not receive widespread attention among the statisticians, and Haldane (1948) rediscovered the same concept. He introduced the term 'geometric median' to distinguish it from 'arithmetic median', which is the vector of coordinatewise medians. Gower (1974) referred to it as 'mediancentre' while discussing an algorithm for computing it. Brown (1983) introduced the term 'spatial median', which so far has been most popular. But none of these multivariate analogs of medians are equivariant under general affine transformations. Tukey (1975) defined another notion of multivariate median, called the 'half-space depth median' based on the geometry of the data cloud. Similarly, Oja (1983) and Liu (1990) proposed 'simplicial or generalized median' and 'simplicial depth median' respectively, which are also based on the geometry of the data cloud. These three notions of multivariate median are equivariant under affine transformations, but all of them are computationally quite complex. In the later sections, we will discuss these popular notions of multivariate medians in little detail with their advantages and shortcomings.

1.2 Coordinatewise Median and Coordinatewise Sign Test

Most simple and oldest notion of multivariate median is the vector of coordinatewise medians. We may define it formally as follows.

Definition 1.2.1 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be d-dimensional observations. We define the vector of coordinatewise median to this data to be any point $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^d$, which minimizes $\sum_{i=1}^{n} |\boldsymbol{X}_i - \boldsymbol{\theta}|$, where $|\boldsymbol{x}| = |x_1| + |x_2| + \cdots + |x_d|$ for $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$.*

This multivariate median is easy to compute as it involves computing univariate medians only. It may be noted that this median vector has a breakdown point of 50%, which was established by Lopuhaa and Rousseeuw (1991). Bickel (1964) derived the asymptotic normality and $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_n$. At the same time, he pointed out that this estimator loses efficiency compared to sample mean in the presence of high correlations among the coordinates of the data vectors. He commented that this pathological behaviour of the estimate may be due to lack of affine equivariance of the proposed estimate. We will discuss in detail this issue of efficiency in Chapter 2 in relation with our proposed estimate of location.

Based on this vector of coordinatewise medians, Bickel (1965) constructed a coordinatewise sign test statistic as an alternative to Hotelling's T^2 for the multivariate location parameter. He established the asymptotic normality of the proposed statistic and suggested a suitable chi-square test. The main advantage of this method is that it is very easy to calculate the test statistic and its asymptotic distribution is normal. But these coordinatewise procedures are handicapped by the fact that they are not equivariant (or, not invariant in the case of test statistic) under arbitrary affine transformations (not even under rotations) though they are equivariant under location shift and coordinatewise scale transformations.

1.3 Spatial Median and Angle Test

Weber (1909) considered a problem in the 'location theory' in which a company has to select an appropriate location for its warehouse, which will serve n customers whose planar coordinates are given by $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$. He assumed that the company can locate the warehouse at any coordinate without any constraint, and the transportation cost for deliveries from the warehouse to the customers are proportional to Euclidean distances only. As a solution to this location problem, Weber (1909) suggested to minimize the average transportation cost (or total transportation cost) from the warehouse to the customers. Gini

and Galvani (1929) introduced this concept as a new definition of multivariate median. Haldane (1948) coined the term 'geometric median' which is defined as follows :

Definition 1.3.1 *For d-dimensional observation vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, the spatial median or geometric median is defined as the point $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^d$, which minimizes the sum of Euclidean distances of the data points from it, i.e.*

$$\hat{\boldsymbol{\theta}}_n = arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{\theta}\|,$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_d^2}$, $\boldsymbol{x} = (x_1, \ldots, x_d)^T$.

Note that spatial median reduces to standard univariate median for $d = 1$, and as the Euclidean distance is invariant under rotations or orthogonal transformations, spatial median is automatically equivariant under rotations. Apart from the fact that it is not equivariant under arbitrary affine transformations, it is not equivariant under coordinate-wise scale transformations either. Thus, if different coordinates of the data vectors are measured in different scales, spatial median does not lead to any meaningful estimate of location. Despite its limitations, spatial median leads to quite efficient estimates in spherically symmetric models, and its efficiency increases with the dimension of the observations (see Brown 1983, Chaudhuri 1992a). Besides, good algorithms are available for computing spatial median (Gower 1974, Bedall and Zimmermann 1979 etc.).

An interesting geometrical fact is that the gradient vectors $(\boldsymbol{X}_i - \boldsymbol{\theta})/\|\boldsymbol{X}_i - \boldsymbol{\theta}\|$ are uniformly distributed over d-dimensional unit sphere if the common distribution of $\boldsymbol{X}_i$'s are spherically symmetric with $\boldsymbol{\theta}$ as the point of spherical symmetry. Thus, a test [some times called the "angle test", see Brown (1983)] for the center of spherical symmetry of the data vectors may be constructed based on these direction vectors, and the test will reduce to a test of uniformity on the sphere. Mardia (1972) discussed the problem in the context of directional statistics. In Chapter 2, we will discuss an affine equivariant version of spatial median and the related invariant version of angle test for locations will be discussed in Chapter 3.

Chaudhuri (1992a) and Neimiro (1992) had almost simultaneously studied the asymptotic distribution of spatial median under very general conditions and using different approaches. They showed that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges asymptotically to a d-dimensional normal distribution. Chaudhuri (1992a) has also showed that the efficiency of the spatial median over sample mean increases to one as the dimension $d \longrightarrow \infty$ when the underlying distribution is multivariate spherical normal. Lopuhaa and Rousseeuw (1991) have investigated the breakdown of the spatial median and have found it to be 50%, which is same as that of univariate median.

1.4 Oja's Simplicial Median and Related Sign Test

Oja (1983) proposed an interesting notion of affine equivariant multivariate median. Consider a point $\theta \in \mathbb{R}^d$, which is a point of central or elliptic symmetry of a distribution F. In univariate problems, the absolute distance $|X_i - \theta|$ can be interpreted as the length of a simplex with vertices X_i and θ, and median is well known as a minimizer of the sum of lengths of all such simplices. Given a sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$, let us define $V[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_d, \theta]$ to be the d-dimensional volume of the simplex in $\mathbb{R}^d$ whose vertices are $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \theta$.

Definition 1.4.1 *Oja's simplicial median of the data set $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a point $\hat{\theta}$ which minimizes*

$$\sum_{1 \le i_1 < \cdots < i_d \le n} V[\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_d}, \theta],$$

where the sum is taken over all subsets of integers $\{i_1, \ldots, i_d\} \subseteq \{1, \ldots, n\}$.

It can be easily verified that Oja's simplicial median does not have the uniqueness property of the spatial median but has the advantage of being affine equivariant. It has a geometric interpretation via gradient vectors. For each $(d-1)$-dimensional simplex with vertices $\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_d}$ chosen as in Definition 1.4.1, construct a vector $A_{i_1 \ldots i_d}(\theta)$ at the origin whose length is proportional to the $(d-1)$-dimensional volume of the simplex, and pointing in the same direction as the ray from θ which passes perpendicularly through the $(d-1)$-dimensional hyperplane generated by the simplex. If the vector sum of all these $\binom{n}{d}$ vectors is zero, then θ is an Oja's simplicial median. See Brown and Hettmansperger (1987, 1989) for the development of this idea and in particular for the construction of affine invariant analogs of rank tests in one and two sample multivariate settings.

Arcones, Chen and Giné (1994) have established the asymptotic normality of the simplicial median under very general conditions through U-statistics type representations. But computation of Oja's simplicial median involves optimization using simplex methods that becomes difficult for large dimensions [see Niinimaa, Oja and Nyblom (1992)]. Also, due to the computational complexity, it is virtually impossible to have an estimate of finite sample variation of the estimate. Niinimaa and Oja (1995) showed that the influence function of the simplicial median is bounded but Oja, Niinimaa and Tableman (1990) have found that this median has 0% breakdown.

As we have noted earlier that the gradient vector of Oja's criterion function is zero at Oja's simplicial median. This implies that one can construct a multivariate analog of sign test based on this gradient vector. In other words, under the null hypothesis of $H_0 : \theta = 0$, origin should be close to the simplicial median computed from the data. Brown

and Hettmansperger (1989) studied the test in bivariate case and derived its asymptotic properties. Hettmansperger, Nyblom and Oja (1994) had shown that the permutation test based on this statistic has an asymptotic χ^2 distribution. They have noted that in the spherically symmetric case this test statistic and angle test as well as Randle's (1989) sign test statistic yield the same asymptotic test with the same asymptotic efficiency.

1.5 Tukey's Halfspace Median and Hodges's Sign Test

Hotelling (1929) introduced a different interpretation of the univariate median which generalizes to what is called the halfspace median due to Tukey (1975) in higher dimensions. It is easy to see that $\min[F(x), 1 - F(x-)]$ is maximized when x is the median of the distribution F. The symmetrized quantile $\min[F(x), 1 - F(x-)]$ is a measure of depth of the point x within the distribution F. If F happens to be the empirical distribution of a sample $X_1, \ldots, X_n$, the above leads to a minimax interpretation of sample median. The extension of this argument in higher dimensions is straightforward. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a sample in $\mathbb{R}^d$. Let $\mathcal{H}$ be the class of all closed halfspaces in $\mathbb{R}^d$. Following Tukey (1975), we define the halfspace depth $HD(\boldsymbol{\theta})$ of a point $\boldsymbol{\theta} \in \mathbb{R}^d$ within the data set as

$$HD(\boldsymbol{\theta}) = n^{-1} \inf \left\{ \sum_{i=1}^{n} 1\{\boldsymbol{X}_i \in H\} \right\},$$

where the infimum is taken over all closed halfspaces $H \in \mathcal{H}$ for which $\boldsymbol{\theta} \in H$.

Definition 1.5.1 *The halfspace median of a data set is defined as*

$$\hat{\boldsymbol{\theta}} = arg \max_{\boldsymbol{\theta} \in \mathbb{R}^d} HD(\boldsymbol{\theta}).$$

The breakdown properties of the halfspace median have been extensively studied by Donoho and Gasko (1992) and Chen (1995). In particular, the breakdown point of the halfspace median is at least $1/(d+1)$ and as high as $1/3$ in the limit for large samples from a centrally symmetric distribution. This median is affine equivariant. Though geometrically appealing, the asymptotic properties of it is yet to be fully worked out, and it is computationally quite complex. Ruts and Rousseeuw (1996) developed an algorithm to compute the halfspace depth and halfspace median for bivariate data. For dimensions $d \geq 3$, there is not yet any good algorithm for computing this median.

Chaudhuri and Sengupta (1993) have shown that the multivariate extension of Hodges bivariate sign test statistic (Hodges, 1955) is equivalent to the halfspace depth of the origin within the data cloud in a d-dimensional Euclidean space, and this test uses it to make the decision about the null hypothesis that asserts the origin as a halfspace median of the probability distribution generating the data.

1.6 Liu's Simplicial Depth Median and Related Sign Test

An affine equivariant multivariate median and affine invariant notion of data depth can be obtained from the simplicial depth function of Liu (1990). To motivate it, we observe that the usual sample median in one dimension can be characterized by the fact that it lies in the greatest number of intervals constructed from the data points. In this sense, it can be viewed as being deep inside the data cloud. To generalize these ideas in higher dimensions, it suffices to replace intervals by d-dimensional simplices in $\mathbb{R}^d$. The empirical depth function is defined to be

$$SD(\boldsymbol{\theta}) = \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < \cdots < i_{d+1} \le n} 1\{\boldsymbol{\theta} \in S(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d+1}})\}$$

where $S(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1})$ is the simplex with vertices $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}$.

Definition 1.6.1 *A simplicial depth median is a point $\hat{\boldsymbol{\theta}}$ which minimizes the function $SD(\boldsymbol{\theta})$ over all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

It is evident that this depth function is affine invariant, and the corresponding simplicial depth median is equivariant under arbitrary affine transformations. Arcones, Chen and Giné (1994) have established the asymptotic normality of this median under very general conditions using U-statistics type representations. Rousseeuw and Ruts (1996) have developed an algorithm to compute the simplicial depth of a point and the simplicial depth median for bivariate data. However, it is computationally quite complex in dimensions $d \ge 3$ and no good algorithm is yet available. Due to computational complexity, it is nearly impossible to estimate finite sample variations of this median even in dimension $d = 2$. Chen (1996) has shown that it has a breakdown point less than $1/(d+1)$.

Chaudhuri and Sengupta (1993) have shown that for $d = 2$, Liu's simplicial depth function is related in a very interesting way to a statistic used by Ajne (1968) (see also Oja and Nyblom, 1989) for testing the uniformity of a circular distribution. It is a matter of simple and straight forward algebra to verify that in the bivariate case, Ajne's test statistic is equivalent to the statistic that counts the number of triangles (simplices in two-dimensional Euclidean space), which are formed with the data points as their vertices and contain the origin in $\mathbb{R}^2$ as an interior point. In other words, it is the simplicial depth of the origin in $\mathbb{R}^2$. Unfortunately, an analogous result fails to hold in any of the dimensions $d \ge 3$.

1.7 A New Approach

From the discussions in the earlier sections it is clear that computationally simpler multivariate medians like the vector of coordinate wise medians and spatial median are not equivariant under arbitrary affine transformations of the data vectors, whereas affine equivariant multivariate medians like Oja's simplicial median, Tukey's halfspace median and Liu's simplicial depth median are computationally quite complex. To resolve this problem, we propose a new approach to construct multidimensional estimates and tests in this thesis. Our approach is based on a transformation retransformation strategy that owes its origin to 'data-driven coordinate system', which was introduced by Chaudhuri and Sengupta (1993). In Chapter 2, we introduce the transformation retransformation strategy and construct several estimates of multivariate location parameter, which are affine equivariant as well as computationally simple. In the process of studying the properties of the proposed estimates, we observe some intriguing facts about the statistical efficiency of multivariate location estimates. The contents of Chapter 2 have been drawn mainly from Chakraborty and Chaudhuri (1996, 1998a, 1998b) and Chakraborty, Chaudhuri and Oja (1998). In Chapter 3, we construct some multivariate analogs of sign and rank tests, which are affine invariant, based on our transformation retransformation strategy. The performance of the tests are studied analytically as well as using some simulations and real data analysis. Chapter 3 is based on Chakraborty and Chaudhuri (1998b) and Chakraborty, Chaudhuri and Oja (1998). In Chapter 4, we discuss the multiresponse linear models and extend our transformation retransformation strategy to construct affine equivariant estimates of the parameter matrices there. We have considered two approaches: one extends the usual least absolute deviation or median regression in multidimension, and the second one extends the rank regression in multidimension. In addition to carrying out a detailed theoretical study, the performance of the proposed methodology has been demonstrated by results from some simulation studies and the analysis of some real data sets. The contents of Chapter 4 are drawn mainly from Chakraborty (1997a) and Chakraborty and Chaudhuri (1997). In Chapter 5, we introduce a new notion of affine equivariant multivariate quantile with the help of transformation retransformation strategy. These multidimensional quantiles are useful in constructing multivariate Q-Q plots and quantile contour plots. These quantiles and the plots open up the possibility of constructing several descriptive statistics for multivariate data clouds. The contents of Chapter 5 are based on Chakraborty (1997b).

Chapter 2

Estimation of Multivariate Location

2.1 Transformation and Retransformation : Methodology and Motivation

Let us begin by observing a simple geometrical fact about any given affine transformation of a set of multivariate observations. For a nonsingular $d \times d$ matrix $\mathbf{A}$ and any $b \in \mathbb{R}^d$, the transformation that maps the d-dimensional observations $\boldsymbol{X}_i$ into $\mathbf{A}\boldsymbol{X}_i + b$ for $1 \leq i \leq n$ essentially expresses the original data in terms of a new coordinate system determined by $\mathbf{A}$ and b. The new origin is located at $-\mathbf{A}^{-1}b$, and depending on whether $\mathbf{A}$ is an orthogonal matrix or not, this new coordinate system may or may not be an orthonormal system. The fundamental idea that lies at the root of data based transformation and retransformation is to form an appropriate 'data driven coordinate system' [see also Chaudhuri and Sengupta (1993)] and to express all the data points in terms of that coordinate system first. After the selection of a 'data driven coordinate system', one computes a location estimate or a test statistic based on those transformed data points. Finally, the location estimate is retransformed to express it back in terms of the original coordinate system. In order to form a 'data driven coordinate system', we need $d+1$ data points in $\mathbb{R}^d$, one of which will determine the origin, and the lines joining that origin to the remaining d data points will form various coordinate axes. To get a valid coordinate system, these $d+1$ points must satisfy some 'nonsingularity' or 'affine independence' condition. However, it is not necessary for this 'data driven coordinate system' to be an orthonormal system. We will now discuss in detail and in more precise terms how this transformation and retransformation technique converts different multivariate location estimates into affine

equivariant estimates of multivariate location.

Consider data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$. Unless specified otherwise, all vectors will be column vectors, and the superscript T will be used to denote the transpose of a vector or a matrix. Define

$$S_n = \{\alpha | \alpha \subseteq \{1, 2, \ldots, n\} \text{ and } \#\{i : i \in \alpha\} = d+1\},$$

which is the collection of all subsets of size $d+1$ of $\{1, 2, \ldots, n\}$. For a fixed $\alpha = \{i_0, i_1, \ldots, i_d\} \in S_n$, let $\mathbf{X}(\alpha)$ be the $d \times d$ matrix whose columns are the random vectors $(\boldsymbol{X}_i - \boldsymbol{X}_{i_0})$ with $i \in \alpha$ and $i \neq i_0$. We assume that the elements of α are naturally ordered, and if the $\boldsymbol{X}_i$'s are independent and identically distributed with a common probability distribution that happens to be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, $\mathbf{X}(\alpha)$ must be an invertible matrix with probability one. We will treat $\mathbf{X}(\alpha)$ as a data based transformation matrix, and for each $i \in \alpha$, write $\boldsymbol{Y}_i^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_i$ (cf. data driven coordinate system discussed in Chaudhuri and Sengupta (1993)). Note that here we are trying to view the data cloud from a data-centric reference frame created by the basis matrix $\mathbf{X}(\alpha)$. Consider

$$\boldsymbol{Z}_i^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_{i_0}) = \boldsymbol{Y}_i^{(\alpha)} - \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_{i_0}.$$

A simple but crucial fact about these tranformed observations can be stated as follows.

Proposition 2.1.1 *Fix an $\alpha \in S_n$, and let the common distribution of the independent and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be absolutely continuous in $\mathbb{R}^d$. Then the transformed vectors $\boldsymbol{Z}_i^{(\alpha)}$'s with $1 \leq i \leq n$ and $i \notin \alpha$ as defined above form a maximal invariant with respect to the group of invertible affine transformations on $\mathbb{R}^d$.*

Proof : Let $\mathbf{A}$ be a $d \times d$ nonsingular matrix and $\boldsymbol{b}$ be a d-dimensional vector and let $\boldsymbol{Y}_i = \mathbf{A}\boldsymbol{X}_i + \boldsymbol{b}$ for $1 \leq i \leq n$. Then it is easy to see that $\mathbf{Y}(\alpha) = \mathbf{A}\mathbf{X}(\alpha)$. Hence $\{\mathbf{Y}(\alpha)\}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Y}_{i_0}) = \{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_{i_0})$ for all $1 \leq i \leq n$. This ensures the invariance of the $\boldsymbol{Z}_i^{(\alpha)}$'s under invertible affine transformations. Also, for two sets of data points $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$, if we have $\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_{i_0}) = \{\mathbf{Y}(\alpha)\}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Y}_{i_0})$ for all $i \notin \alpha$, we automatically have

$$\boldsymbol{X}_i = \mathbf{X}(\alpha)\{\mathbf{Y}(\alpha)\}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Y}_{i_0}) + \boldsymbol{X}_{i_0} \tag{2.1}$$

for all i such that $1 \leq i \leq n$. Note that the equation is trivially true for $i \in \alpha$. Therefore, the data set $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ is obtainable from $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n\}$ by an invertible affine

transformation and vice versa. This completes the proof by estabilishing the maximality of the invariant vectors. □

Now we proceed to demonstrate how the proposed *transformation retransformation* methodology can be used as a general tool to construct affine equivariant estimates of multivariate location out of non-equivariant ones. Let $\hat{\phi}_n^{(\alpha)}$ be some translation invariant estimate of multivariate location based on d-dimensional transformed observations $\boldsymbol{Y}_i^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_i$ such that $1 \leq i \leq n$ and $i \notin \alpha$. Then define the location estimate $\hat{\boldsymbol{\theta}}_n^{(\alpha)}$ for the original data by retransforming $\hat{\phi}_n^{(\alpha)}$ as

$$\hat{\boldsymbol{\theta}}_n^{(\alpha)} = \{\mathbf{X}(\alpha)\}\hat{\phi}_n^{(\alpha)}.$$

The following Proposition asserts affine equivariance of $\hat{\boldsymbol{\theta}}_n^{(\alpha)}$.

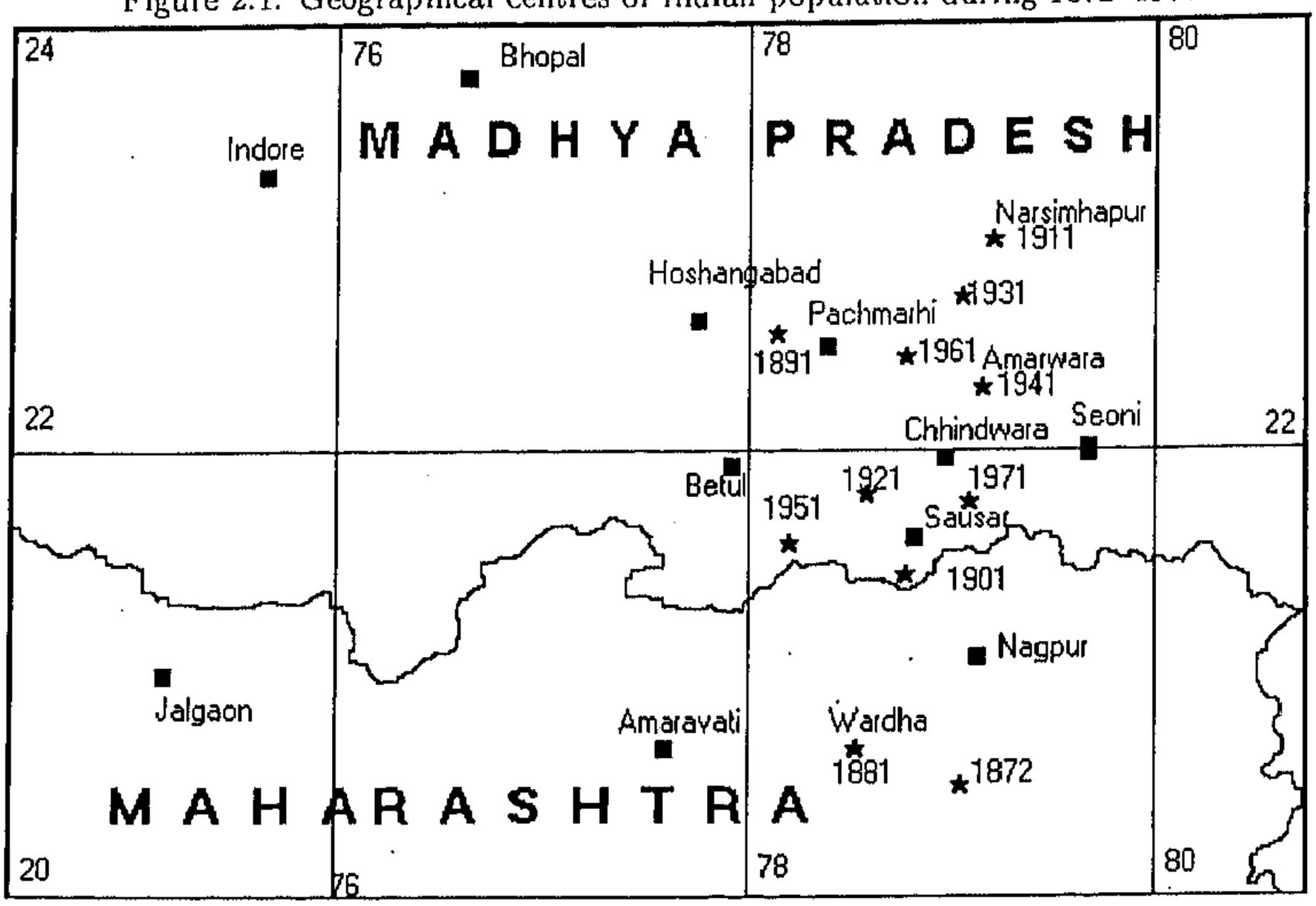
Proposition 2.1.2 *Let $\hat{\boldsymbol{\theta}}_n^{(\alpha)}$ be the location estimate based on the data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as described above. Suppose that $\mathbf{A}$ is a fixed $d \times d$ nonsingular matrix and $\boldsymbol{b}$ is a fixed vector in $\mathbb{R}^d$. Then the multivariate location estimate computed from $\mathbf{A}\boldsymbol{X}_1 + \boldsymbol{b}, \mathbf{A}\boldsymbol{X}_2 + \boldsymbol{b}, \ldots, \mathbf{A}\boldsymbol{X}_n + \boldsymbol{b}$ in the same way as above (using the same index set α) will be $\mathbf{A}\hat{\boldsymbol{\theta}}_n^{(\alpha)} + \boldsymbol{b}$.*

Proof : First observe that in view of the way the matrix $\mathbf{X}(\alpha)$ has been constructed, if the $\boldsymbol{X}_i$'s are transformed to $(\mathbf{A}\boldsymbol{X}_i + \boldsymbol{b})$'s, $\mathbf{X}(\alpha)$ will be transformed to $\mathbf{A}\mathbf{X}(\alpha)$. Also, note that the $\boldsymbol{Y}_i^{(\alpha)}$'s remain invariant under a linear transformation of the $\boldsymbol{X}_i$'s. Hence, in view of the location equivariance of $\hat{\phi}_n^{(\alpha)}$, for the transformed data points $(\mathbf{A}\boldsymbol{X}_i + \boldsymbol{b})$, $\hat{\phi}_n^{(\alpha)}$ will be transformed to $\hat{\phi}_n^{(\alpha)} + \{\mathbf{A}\mathbf{X}(\alpha)\}^{-1}\boldsymbol{b}$. Consequently $\hat{\boldsymbol{\theta}}_n^{(\alpha)}$, which was defined as $\{\mathbf{X}(\alpha)\}\hat{\phi}_n^{(\alpha)}$, will be transformed to $\mathbf{A}\hat{\boldsymbol{\theta}}_n^{(\alpha)} + \boldsymbol{b}$. □

Before we discuss the asymptotics and other properties of the estimate $\hat{\boldsymbol{\theta}}_n^{(\alpha)}$ when $\hat{\phi}_n^{(\alpha)}$ is some specific estimate of multivariate location, e.g. vector of coordinatewise medians, spatial median, vector of coordinatewise Hodges-Lehmann estimate etc. in the following sections, let us now consider the following example as an illustration of the methodology, where we will try to locate the 'geographical centre' of Indian population using transformation retransformation coordinatewise median computed from decennial census data. This example will demonstrate the usefulness of this affine equivariant location estimate as a multivariate descriptive statistic.

Example 2.1 : To estimate the 'geographical centre' of a population distribution, earlier statisticians used centroid (i.e. usual multivariate mean) but observed that the centroid may be highly sensitive to the influence of probability masses at the extremes [see e.g. Small (1990) and Chaudhuri (1996)]. In other words, an event like a death or a birth in the

Figure 2.1: Geographical centres of Indian population during 1872–1971



* indicates a population centre.

periphery of the country tends to have more influence on the centroid of the population than a similar event occurring at the central part of the country. This motivates the use of a median like measure of the centre of a population. For India, we have used the data obtained in census years during the period 1872 to 1971 and considered only the populations of Type-I towns (as classified in 1971), which cover nearly 80% of the population. The rest of the population is scattered in smaller towns and villages, which have insignificant effect on the estimation of the centre of the population, and by ignoring them we have substantially reduced the time required for the compilation of the data and subsequent numerical computation. As the radius of the earth is very large compared to the size of India, we have ignored the effect of the curvature of the earth in this example, and the population is regarded as living on an essentially flat map in which the lines of lattitude and those of longitude are assumed to be orthogonal straight lines. The 'geographical centres' of population as located by our transformation retransformation coordinatewise median are given in Figure 2.1.

2.2 Vector of Coordinatewise Medians

Let us now consider $\hat{\phi}_n^{(\alpha)}$ to be the vector of coordinatewise medians of the transformed observations $\boldsymbol{Y}_i^{(\alpha)}$, with $1 \leq i \leq n$ and $i \notin \alpha$. Define the resulting transformation retransformation multivariate median as $\hat{\theta}_M^{(\alpha)} = \{\mathbf{X}(\alpha)\}\hat{\phi}_n^{(\alpha)}$. It is easy to see that

$$\hat{\theta}_M^{(\alpha)} = \arg \min_{\theta \in \mathbf{R}^d} \sum_{i \notin \alpha} |\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \theta)|$$

where $|\cdot|$ is usual l_1-norm, i.e. $|x| = |x_1| + \cdots + |x_d|$ for $x = (x_1, \ldots, x_d)^T$.

In the case of data arising from an elliptically symmetric distribution, $\hat{\theta}_M^{(\alpha)}$ estimates the centre of elliptic symmetry of that distribution. In general, $\hat{\theta}_M^{(\alpha)}$ can be viewed as a descriptive statistic that yields a new concept of location of a multivariate data cloud. Note at this point that for $d = 1$ and a fixed $\alpha = \{i_0, i_1\} \in S_n$, $\hat{\theta}_M^{(\alpha)}$ reduces to the usual univariate median of the X_i's excluding the observations X_{i_0} and X_{i_1}. Hence the difference between $\hat{\theta}_M^{(\alpha)}$ and the median of all X_i's with $1 \leq i \leq n$ will be insignificant especially when the sample size n is large, and their asymptotic behaviour will be identical. Specifically, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed univariate observations with a common density f that has a median at θ, and f is continuous and positive at θ, the asymptotic distribution of $\sqrt{n}(\hat{\theta}_M^{(\alpha)} - \theta)$ will be Gaussian with mean 0 and variance $\{2f(\theta)\}^{-2}$, which is the same as the asymptotic distribution of the median of all the X_i's (see Bahadur, 1966). The situation however is very different in higher dimensions. For $d \geq 2$, the asymptotic behaviour of $\hat{\theta}_M^{(\alpha)}$ critically depends on the matrix $\mathbf{X}(\alpha)$, and as we will see later, the selection of $\alpha \in S_n$ and $i_0 \in \alpha$ has a crucial impact on the asymptotic performance of $\hat{\theta}_M^{(\alpha)}$.

It is worthwhile to note at this stage that though transforming the data points by the square root of the sample variance covariance matrix is a popular approach, the resulting coordinate system does not have any simple and natural geometric interpretation. Further, such a transformation cannot lead to an affine equivariant modification of nonequivariant vector of coordinatewise median, and the limitation of that approach is primarily due to the fact that there does not exist an affine equivariant square root of the variance covariance matrix. On the other hand, for a fixed $\alpha \in S_n$, multiplication of the data points with the matrix $\{\mathbf{X}(\alpha)\}^{-1}$ can be viewed as a different (and somewhat unconventional) way of normalizing the observations. Clearly, once we select $\alpha \in S_n$, the computation of $\hat{\phi}_n^{(\alpha)}$ and $\hat{\theta}_M^{(\alpha)}$ is extremely simple in any dimension. One only needs to compute the usual univariate median for each coordinate of the transformed observations $\boldsymbol{Y}_i^{(\alpha)}$, and then retransform the resulting vector of medians (i.e. $\hat{\phi}_n^{(\alpha)}$) by multiplying it with $\mathbf{X}(\alpha)$. We now state the following Theorem, which exposes an interesting geometric feature of $\hat{\theta}_M^{(\alpha)}$.

Theorem 2.2.1 *Fix $\alpha \in S_n$ and $i_0 \in \alpha$, and let $\theta \in \mathbb{R}^d$. For each $i \in \alpha$, replace $\boldsymbol{X}_i$ by $\boldsymbol{Z}_i = \boldsymbol{X}_i + \theta$, and for each $i \notin \alpha$, replace $\boldsymbol{X}_i$ by $\boldsymbol{Z}_i = \boldsymbol{X}_i + \boldsymbol{X}_{i_0}$. So, each data point is transformed by a location shift, where the shifting vector is either θ or $\boldsymbol{X}_{i_0}$ depending on whether the data point to be shifted is used in the formation of the transformation matrix $\mathbf{X}(\alpha)$ or not, respectively. Consider those simplices in $\mathbb{R}^d$ each of which is formed by a collection of $d+1$ points $\{\boldsymbol{Z}_{j_1}, \boldsymbol{Z}_{j_2}, \ldots, \boldsymbol{Z}_{j_{d-1}}, \boldsymbol{Z}_{i_0}, \boldsymbol{Z}_i\}$ such that $\{j_1, j_2, \ldots, j_{d-1}\} \subset \alpha$ and $i \notin \alpha$. Then $\theta = \hat{\theta}_M^{(\alpha)}$ minimizes the sum of volumes of all such simplices.*

Proof : For $z = (z_1, z_2, \ldots, z_d) \in \mathbb{R}^d$, let $|z|$ denote the l_1-norm of z defined as $z = \sum_{i=1}^d |z_i|$. Then as we have noted earlier

$$\hat{\theta}_M^{(\alpha)} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{1 \le i \le n \,;\, i \notin \alpha} |\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \theta)|.$$

Now $\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \theta)$ can be viewed as a solution (in z) of the system of linear equations $\{\mathbf{X}(\alpha)\}z = (\boldsymbol{X}_i - \theta)$. So, if one applies the well-known Cramer's rule for solving a system of linear equations, the absolute value of any component of the d-dimensional vector $\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \theta)$ will be of the form

$$|\det\{\mathbf{X}(\alpha)\}|^{-1} |\det\{(\boldsymbol{X}_{j_1} - \boldsymbol{X}_{i_0}, \boldsymbol{X}_{j_2} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{j_{d-1}} - \boldsymbol{X}_{i_0}, \boldsymbol{X}_i - \theta)\}|.$$

The proof of the Theorem is now complete in view of the fact that

$$|\det\{(\boldsymbol{X}_{j_1} - \boldsymbol{X}_{i_0}, \boldsymbol{X}_{j_2} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{j_{d-1}} - \boldsymbol{X}_{i_0}, \boldsymbol{X}_i - \theta)\}|$$

is the volume of the simplex in $\mathbb{R}^d$, which is formed by the collection of $d+1$ points $\{\boldsymbol{Z}_{j_1}, \boldsymbol{Z}_{j_2}, \ldots, \boldsymbol{Z}_{j_{d-1}}, \boldsymbol{Z}_{i_0}, \boldsymbol{Z}_i\}$ as described in the statement of the Theorem. □

Random simplices formed by data points play a very crucial role in the construction of Oja's median (1983) as well as Liu's median (1990) as we have described in Chapter 1. The above Theorem indicates that they have a fundamental role in the construction of $\hat{\theta}_M^{(\alpha)}$ too.

2.2.1 Asymptotic Properties of Proposed Median

From now on we will assume that the $\boldsymbol{X}_i$'s are independent and identically distributed observations with a common probability distribution that is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Then, for a fixed $\alpha \in S_n$, the transformed observations $\boldsymbol{Y}_i^{(\alpha)}$'s with $1 \le i \le n$ and $i \notin \alpha$ are conditionally independently distributed with a common absolutely continuous distribution if we condition on the $\boldsymbol{X}_i$'s for which $i \in \alpha$. It

is now obvious that the limiting conditional distributions (conditioned on the $\boldsymbol{X}_i$'s with $i \in \alpha$) of both $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ and $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$ will be normal in view of well-known asymptotic results about the univariate median that are applicable to a vector of univariate medians (see e.g. Babu and Rao, 1988). When the common distribution of the $\boldsymbol{X}_i$'s happens to be elliptically symmetric, it is possible to describe that limiting normal distribution explicitly by deriving a useful expression for the limiting dispersion matrix in terms of $\mathbf{X}(\alpha)$. This leads to valuable insights into the asymptotic performance of $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$ as an estimate of the center of elliptic symmetry, and provides us with a way of adaptively selecting an optimal $\alpha \in S_n$ for forming the transformation matrix $\mathbf{X}(\alpha)$.

2.2.1.1 Behaviour in the Elliptically Symmetric Case

Suppose that the $\boldsymbol{X}_i$'s have an elliptically symmetric probability distribution with density $\{\det(\Sigma)\}^{-1/2} f\{(x-\theta)^T\Sigma^{-1}(x-\theta)\}$. Here $\boldsymbol{\theta} \in \mathbb{R}^d$ is the location of symmetry, and Σ is a $d \times d$ positive definite matrix. Let us write $\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$, where $\mathbf{R}(\alpha)$ is a diagonal matrix with positive diagonal entries and $\mathbf{J}(\alpha)$ is a matrix whose rows are of unit length. Clearly, the rows of $\mathbf{J}(\alpha)$ are obtained by normalizing the rows of $\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}^{-1}$, and the diagonal elements of $\mathbf{R}(\alpha)$ are the norms of those rows.

Theorem 2.2.2 *Fix $\alpha \in S_n$ and $i_0 \in \alpha$ as before. Assume that the density function f is such that any univariate marginal g of the spherically symmetric density $f(x^Tx)$ is differentiable and positive at zero. Then as n tends to infinity, the conditional distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_M^{(\alpha)} - \boldsymbol{\theta})$ given the $\boldsymbol{X}_i$'s with $i \in \alpha$ converges weakly to a d-dimensional normal distribution with zero mean and the dispersion marix $c\Sigma^{1/2}\{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^T\}^{-1}\Sigma^{1/2}$. Here $c = \{2g(0)\}^{-2}$, and $\mathbf{D}(\alpha)$ is the $d \times d$ matrix whose diagonal elements are all equal to 1, and for $i \neq j$, its (i,j)-th element is $(2/\pi)\sin^{-1}\gamma_{ij}$, γ_{ij} being the inner product of the i-th and the j-th rows of $\mathbf{J}(\alpha)$.*

Proof : In view of affine equivariance of the location estimate $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$, it is sufficient to prove the Theorem in the special case when $\boldsymbol{\theta}$ is the zero vector in $\mathbb{R}^d$ and Σ is the $d \times d$ identity matrix. Then, given the $\boldsymbol{X}_i$'s for which $i \in \alpha$, the transformed observations $\boldsymbol{Y}_i^{(\alpha)}$'s with $i \notin \alpha$ are conditionally i.i.d random vectors with common density $|\det\{\mathbf{X}(\alpha)\}| f\{y^T[\mathbf{X}(\alpha)]^T[\mathbf{X}(\alpha)]y\}$. Let $r_1, \ldots, r_d$ be the diagonal entries of $\mathbf{R}(\alpha)$. In view of the main result in Babu and Rao (1988) on asymptotic distribution of the vector of univariate quantiles of a multivariate data, the conditional distribution of $n^{1/2}\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ will converge weakly to a d-variate normal distribution with zero mean, and the limiting dispersion matrix will be such that its k-th diagonal entry will be cr_k^2, and for $k \neq l$,

its (k,l)-th element will be $4cr_kr_l\{Pr(\ U_{ik}^{(\alpha)} > 0 \text{ and } U_{il}^{(\alpha)} > 0\ ) - 1/4\}$, where c is as defined in the Theorem. Here $U_{ik}^{(\alpha)}$ and $U_{il}^{(\alpha)}$ are the k-th and the l-th components of $Y_i^{(\alpha)}$ respectively. Note that we are using the fact that for a d-dimensional random vector $\boldsymbol{Z}$ with a spherically symmetric distribution, the distribution of the random variable $a^T\boldsymbol{Z}$ is the same for any $\boldsymbol{a} \in \mathbb{R}^d$ such that $a^Ta = 1$. Also, since the conditional distribution of $Y_i^{(\alpha)}$ is elliptically symmetric around the origin in $\mathbb{R}^d$, $Pr\{\ U_{ik}^{(\alpha)} > 0 \text{ and } U_{il}^{(\alpha)} > 0\ \}$ does not depend on the density f. Recall that the rows of $\mathbf{J}(\alpha)$ are of unit length obtained by normalizing the rows of $\{\mathbf{X}(\alpha)\}^{-1}$. We now have the following by some routine analytic computation.

$$Pr(\ U_{ik}^{(\alpha)} > 0 \text{ and } U_{il}^{(\alpha)} > 0\ ) = 1/4 + (1/2\pi)\sin^{-1}\gamma_{kl}\ .$$

So, the dispersion matrix of the conditional asymptotic distribution of $n^{1/2}\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ is

$$c\{\mathbf{R}(\alpha)\}\{\mathbf{D}(\alpha)\}\{\mathbf{R}(\alpha)\}.$$

Next recall that

$$\hat{\boldsymbol{\theta}}_M^{(\alpha)} = \mathbf{X}(\alpha)\hat{\boldsymbol{\phi}}_n^{(\alpha)} = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{R}(\alpha)\}^{-1}\hat{\boldsymbol{\phi}}_n^{(\alpha)}\ .$$

The proof of the Theorem is now complete by straight-forward algebra. □

It follows from the preceding Theorem that $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$ is a $n^{1/2}$-consistent estimate of $\boldsymbol{\theta}$, and its conditional asymptotic generalized variance is

$$(c/n)^d\{\det(\Sigma)\}[\det\{\mathbf{D}(\alpha)\}][\det\{\mathbf{J}(\alpha)\}]^{-2}\ .$$

Consider now the symmetric positive definite matrix

$$\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^T\}^{-1}\ .$$

Note that it depends only on the directions of the rows of $\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}^{-1}$ and not on their magnitudes. The following Theorem establishes a lower bound for $\det\{\mathbf{V}(\alpha)\}$, and this yields a lower bound for conditional asymptotic generalized variance of $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$.

Theorem 2.2.3 *For the matrices* $\mathbf{D}(\alpha)$ *and* $\mathbf{J}(\alpha)$ *defined above, we have* $\det\{\mathbf{D}(\alpha)\} \geq [\det\{\mathbf{J}(\alpha)\}]^2$ *so that* $\det\{\mathbf{V}(\alpha)\} \geq 1$. *This lower bound is sharp in the sense that an exact equality in place of the inequality will hold if* $\mathbf{J}(\alpha)$ *happens to be an orthogonal matrix.*

The following well-known Fact will be used in the proof of Theorem 2.2.3. A proof of this Fact has been discussed in Lancaster (1969).

Fact 2.2.4 *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be p and q-dimensional normal random vectors. Then*

$$\max_{\boldsymbol{a}\in\mathbb{R}^p,\boldsymbol{b}\in\mathbb{R}^q} Corr(\boldsymbol{a}^T\boldsymbol{X},\boldsymbol{b}^T\boldsymbol{Y}) = \max_{\eta,\psi} Corr(\eta(\boldsymbol{X}),\psi(\boldsymbol{Y}))\,,$$

where "Corr" stands for the usual correlation coefficient, and $\eta : \mathbb{R}^p \to \mathbb{R}$, $\psi : \mathbb{R}^q \to \mathbb{R}$ are measurable functions such that $\eta(\boldsymbol{X})$ and $\psi(\boldsymbol{Y})$ have finite second moments.

Proof of Theorem 2.2.3 : Denote $\{\mathrm{J}(\alpha)\}\{\mathrm{J}(\alpha)\}^T$ by $\mathbf{P}$, and for notational convenience, we will write simply $\mathbf{D}$ for $\mathbf{D}(\alpha)$. We will prove the result by induction on the dimension d. For $d = 2$, the matrices $\mathbf{D}$ and $\mathbf{P}$ can be written as,

$$\mathbf{D} = \begin{bmatrix} 1 & \frac{2}{\pi}\sin^{-1}(\cos\delta) \\ \frac{2}{\pi}\sin^{-1}(\cos\delta) & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 1 & \cos\delta \\ \cos\delta & 1 \end{bmatrix},$$

where $0 < \delta < \pi$ is the angle between the two rows of $\mathrm{J}(\alpha)$. So, $\det(\mathbf{D}) = 1 - \{\frac{2}{\pi}\sin^{-1}(\cos\delta)\}^2 = 1 - (1 - \frac{2}{\pi}\delta)^2$ and $\det(\mathbf{P}) = 1 - \cos^2\delta$. Now, for $0 < \delta < \pi$, $\cos^2\delta \geq (1 - \frac{2}{\pi}\delta)^2$, and the equality holds if and only if $\delta = \pi/2$. This proves the result for $d = 2$.

Now assume that the result is true for dimension $d-1 \geq 2$. Partition the $d \times d$ matrix $\mathbf{D}$ as $\begin{bmatrix} 1 & \boldsymbol{d}^T \\ \boldsymbol{d} & \mathbf{D}_* \end{bmatrix}$ and the $d \times d$ matrix $\mathbf{P}$ as $\begin{bmatrix} 1 & \boldsymbol{p}^T \\ \boldsymbol{p} & \mathbf{P}_* \end{bmatrix}$. Note that $\mathbf{P}$ can be viewed as the correlation matrix of a d-dimensional normal random vector $(U, W_1, \ldots, W_{d-1})$, and $\mathbf{D}$ can be viewed as the correlation matrix of the random vector $(I(U > 0), I(W_1 > 0), \ldots, I(W_{d-1} > 0)) = (V, Z_1, \ldots, Z_{d-1})$ (say), where I is the usual 0–1 valued indicator function. Write $\boldsymbol{W} = (W_1, \ldots, W_{d-1})$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_{d-1})$. Then using Fact 2.2.4 stated above, we get

$$\max_{\boldsymbol{b}\in\mathbb{R}^{d-1}} Corr(U, \boldsymbol{b}^T\boldsymbol{W}) \geq \max_{\boldsymbol{b}\in\mathbb{R}^{d-1}} Corr(V, \boldsymbol{b}^T\boldsymbol{Z})$$

But on the LHS above, we have the multiple correlation coefficient between U and $\boldsymbol{W}$, and on the RHS, we have the multiple correlation coefficient between V and $\boldsymbol{Z}$. Therefore, we must have $\boldsymbol{p}^T\mathbf{P}_*^{-1}\boldsymbol{p} \geq \boldsymbol{d}^T\mathbf{D}_*^{-1}\boldsymbol{d}$. The induction hypothesis implies that $\det(\mathbf{D}_*) \geq \det(\mathbf{P}_*)$. The proof of the Theorem is now complete by observing that $\det(\mathbf{D}) = \{\det(\mathbf{D}_*)\}(1 - \boldsymbol{d}^T\mathbf{D}_*^{-1}\boldsymbol{d})$ and $\det(\mathbf{P}) = \{\det(\mathbf{P}_*)\}(1 - \boldsymbol{p}^T\mathbf{P}_*^{-1}\boldsymbol{p})$. □

2.2.1.2 Adaptive Choice of α

Theorem 2.2.3 implies that whatever may f and Σ be, the conditional asymptotic generalized variance of $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$ cannot be smaller than $(c/n)^d \det(\Sigma)$ for any choice of $\alpha \in S_n$ and $i_0 \in \alpha$, and one should preferably choose $\mathbf{X}(\alpha)$ in such a way that the columns of

$\Sigma^{-1/2}\mathbf{X}(\alpha)$ (i.e. the vectors $\Sigma^{-1/2}(\boldsymbol{X}_i - \boldsymbol{X}_{i_0})$'s, where $i \in \alpha$ and $i \neq i_0$) are as orthogonal as possible so that $\det\{\mathbf{V}(\alpha)\}$ becomes very close to one. Here we propose an adaptive way to select the best subset α. First, obtain some consistent estimate of the scale matrix Σ, say $\hat{\Sigma}$ that is equivariant under nonsingular linear transformation of the data. Then, normalize each data point $\boldsymbol{X}_i$ by $\hat{\Sigma}^{-1/2}$. Define $\boldsymbol{Y}_i = \hat{\Sigma}^{-1/2}\boldsymbol{X}_i$ for $1 \leq i \leq n$. Choose $\alpha \in S_n$ and compute $\hat{v}(\alpha) = \det\{\hat{\mathbf{V}}(\alpha)\}$ based on $\boldsymbol{Y}_i$'s as described before. Then minimize $\hat{v}(\alpha)$ over all choices of $\alpha \in S_n$. Suppose that $\hat{\alpha}$ is the minimizer. Form $\mathbf{X}(\hat{\alpha})$ and compute $\hat{\boldsymbol{\theta}}_M^{(\hat{\alpha})}$ from the original observations $\boldsymbol{X}_i$'s.

Note that once the matrix $\mathbf{X}(\alpha)$ is formed, the computation of $\hat{\boldsymbol{\theta}}_M^{(\alpha)}$ is straightforward as it does not require any further optimization or iterative computation. But the selection of optimal α may require a search over $\binom{n}{d+1}$ possible subsets α, and this number grows very fast with n and d. One can reduce the amount of computation involved for searching the optimal α by stopping whenever $\hat{v}(\alpha)$ is sufficiently close to one in case of transformation retransformation coordinatewise median because we know from Theorem 2.2.3 that the lower bound for $v(\alpha)$ is one. We have observed that this approximation makes the algorithm very fast without making any serious change in the sampling variation or any significant loss of efficiency of the resulting estimate. In all the real examples that we have considered in Section 2.2, it performed satisfactorily. An alternative approach would be to make a random search over different subsets α and stopping when $v(\alpha)$ stabilizes in some appropriate sense. Approaches similar to this have been considered in computing least median of squares estimates [see e.g. Rousseeuw and Leroy (1987)].

2.2.2 Asymptotic Optimality of the Proposed Estimate

In this section, we will discuss some efficiency results of the proposed adaptive transformation retransformation estimate. Suppose that $\alpha^* \in S_n$ minimizes $\det\{\mathbf{V}(\alpha)\} = v(\alpha)$, and recall that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are i.i.d. observations with a common density h on $\mathbb{R}^d$, which need not be elliptically symmetric for the time being.

Theorem 2.2.5 *Assume that h satisfies $\int_{\mathbf{R}^d}\{h(y)\}^{d+1}\,dy < \infty$. Then $v(\alpha^*)$ converges to one in probability as n tends to infinity.*

Proof : Assume without loss of generality that Σ is the d-dimensional identity matrix. Consider $\alpha = \{1, 2, \ldots, d+1\}$ and $i_0 = 1$. As the underlying distribution of the $\boldsymbol{X}_i$'s are i.i.d. with density h, the joint p.d.f. of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}$ can be written as $\prod_{i=1}^{d+1} h(x_i)$. Now we make the following transformation of variables

$$\boldsymbol{Y}_1 = \boldsymbol{X}_2 - \boldsymbol{X}_1, \ldots, \boldsymbol{Y}_d = \boldsymbol{X}_{d+1} - \boldsymbol{X}_1, \boldsymbol{Y}_{d+1} = \boldsymbol{X}_1.$$

Then the joint density of $Y_1, \ldots, Y_{d+1}$ is given by $h(y_{d+1}) \prod_{i=1}^{d} h(y_i + y_{d+1})$. Therefore, the joint density of $Y_1, \ldots, Y_d$ at the origin in $\mathbb{R}^{d \times d}$ is $\int_{R^d} \{h(y)\}^{d+1}\, dy$, which is finite and positive by the condition assumed in the statement of the Theorem. This condition further implies that the map

$$(y_1, y_2, \ldots, y_d) \longmapsto \int_{\mathbb{R}^d} h(y) \prod_{i=1}^{d} h(y_i + y)\, dy$$

from $\mathbb{R}^{d \times d}$ to $\mathbb{R}$ is everywhere continuous. Therefore the joint density of $Y_1, \ldots, Y_d$ must remain bounded away from zero in a neighbourhood of $0 \in \mathbb{R}^{d \times d}$. Consequently, the probability of the event that the columns of $\mathbf{X}(\alpha)$ will be nearly orthogonal (and hence $v(\alpha) = \det\{\mathbf{V}(\alpha)\}$ will be very close to 1) is bounded away from zero. In other words, we have for any $\epsilon > 0$,

$$Pr[\det\{\mathbf{V}(\alpha)\} = v(\alpha) < 1 + \epsilon] \;=\; p_\epsilon > 0$$

Let $\alpha_1, \alpha_2, \ldots, \alpha_{k_n}$ be disjoint subsets of S_n such that k_n tends to infinity as n tends to infinity (e.g. k_n may be equal to $n/(d+1)$). Then

$$\begin{aligned} Pr\{v(\alpha^*) \geq 1 + \epsilon\} &= Pr\{\forall \alpha \in S_n, \quad v(\alpha) \geq 1 + \epsilon\} \\ &\leq Pr\{v(\alpha_1) \geq 1 + \epsilon, \ldots, v(\alpha_{k_n}) \geq 1 + \epsilon\} \\ &= (1 - p_\epsilon)^{k_n} \longrightarrow 0 \;\; as \;\; n \to \infty. \end{aligned}$$

□

Clearly, the integrability condition imposed on h in the above theorem will hold if h happens to be a bounded density on $\mathbb{R}^d$. In the presence of elliptic symmetry with $h(x) = \{det(\Sigma)\}^{-1/2} f\{(x-\theta)^T \Sigma^{-1} (x-\theta)\}$, this condition translates into an integrability condition on f, which is again trivially satisfied for any bounded spherically symmetric density f on $\mathbb{R}^d$. This theorem implies that when the scale matrix Σ is known and the adaptive selection of α^* and S_n is done using that known Σ, the conditional generalized variance of the resulting transformation retransformation estimate tends to the lower bound established in Theorem 2.2.3 in the previous subsection. However, in practice Σ is unknown, and we will estimate it by a consistent and affine equivariant estimate $\hat{\Sigma}$ when we minimize $\hat{v}(\alpha)$ to obtain $\hat{\alpha}$. The next theorem tells that the difference between $v(\hat{\alpha})$ and $v(\alpha^*)$ is asymptotically negligible.

Theorem 2.2.6 *Under the condition assumed in Theorem 2.2.5, $v(\hat{\alpha}) - v(\alpha^*)$ converges in probability to zero as n tends to infinity.*

In order to prove Theorem 2.2.6, we will prove some preliminary results first.

Lemma 2.2.7 $\sup_{\alpha \in S_n} |\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)|$ *converges in probability to zero as* n *tends to infinity, where* $\hat{\mathbf{J}}(\alpha)$ *is obtained in the same way as* $\mathbf{J}(\alpha)$ *using* $\hat{\Sigma}$ *in place of* Σ.

Proof : Let us write $\{\mathbf{X}(\alpha)\}^{-1}\Sigma^{1/2} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$ and similarly $\{\mathbf{X}(\alpha)\}^{-1}\hat{\Sigma}^{1/2} = \hat{\mathbf{R}}(\alpha)\hat{\mathbf{J}}(\alpha)$, where $\hat{\Sigma}$ is a consistent estimate of Σ. Clearly, the rows of $\mathbf{J}(\alpha)$ and $\hat{\mathbf{J}}(\alpha)$ are nothing but the normalized rows of $\{\mathbf{X}(\alpha)\}^{-1}\Sigma^{1/2}$ and $\{\mathbf{X}(\alpha)\}^{-1}\hat{\Sigma}^{1/2}$ respectively. Let the j-th row of $\{\mathbf{X}(\alpha)\}^{-1}$ be u_j^T. Then

$$\frac{u_j^T\hat{\Sigma}^{1/2}}{|u_j^T\hat{\Sigma}^{1/2}|} - \frac{u_j^T\Sigma^{1/2}}{|u_j^T\Sigma^{1/2}|} = \frac{u_j^T\hat{\Sigma}^{1/2}|u_j^T\Sigma^{1/2}| - u_j^T\Sigma^{1/2}|u_j^T\hat{\Sigma}^{1/2}|}{|u_j^T\hat{\Sigma}^{1/2}|\,|u_j^T\Sigma^{1/2}|}$$

$$= \frac{u_j^T(\hat{\Sigma}^{1/2} - \Sigma^{1/2})|u_j^T\Sigma^{1/2}| + u_j^T\Sigma^{1/2}\{|u_j^T\Sigma^{1/2}| - |u_j^T\hat{\Sigma}^{1/2}|\}}{|u_j^T\hat{\Sigma}^{1/2}|\,|u_j^T\Sigma^{1/2}|} .$$

Now, since $\hat{\Sigma} \xrightarrow{P} \Sigma$ (a positive definite matrix) as $n \longrightarrow \infty$, for sufficiently large n and any $d \times 1$ vector u, we must have

$$\frac{u^T\hat{\Sigma}u}{u^Tu} \geq c^2$$

for some $c > 0$. Then

$$\left|\frac{u_j^T\hat{\Sigma}^{1/2}}{|u_j^T\hat{\Sigma}^{1/2}|} - \frac{u_j^T\Sigma^{1/2}}{|u_j^T\Sigma^{1/2}|}\right| \leq \frac{2|\hat{\Sigma}^{1/2} - \Sigma^{1/2}|}{c}$$

Therefore, $\sup_{\alpha \in S_n}\sup_j \left|\frac{u_j^T\hat{\Sigma}^{1/2}}{|u_j^T\hat{\Sigma}^{1/2}|} - \frac{u_j^T\Sigma^{1/2}}{|u_j^T\Sigma^{1/2}|}\right| \leq \frac{2|\hat{\Sigma}^{1/2} - \Sigma^{1/2}|}{c}$. In other words, we must have

$$\sup_{\alpha \in S_n} |\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)| \leq c^*|\hat{\Sigma}^{1/2} - \Sigma^{1/2}| \ ,$$

for some positive constant c^*. The proof is now complete in view of the fact that $\hat{\Sigma}$ is a consistent estimate of Σ. □

Lemma 2.2.8 $\sup_{\alpha \in S_n} |\hat{\mathbf{J}}(\alpha)\{\hat{\mathbf{J}}(\alpha)\}^T - \mathbf{J}(\alpha)\{\mathbf{J}(\alpha)\}^T|$ *converges in probability to zero as* n *tends to infinity.*

Proof : First observe that

$$\begin{aligned}
&|\hat{\mathbf{J}}(\alpha)\{\hat{\mathbf{J}}(\alpha)\}^T - \mathbf{J}(\alpha)\{\mathbf{J}(\alpha)\}^T| \\
&= |\hat{\mathbf{J}}(\alpha)\{\hat{\mathbf{J}}(\alpha)\}^T - \mathbf{J}(\alpha)\{\hat{\mathbf{J}}(\alpha)\}^T \\
&\quad + \mathbf{J}(\alpha)\{\hat{\mathbf{J}}(\alpha)\}^T - \mathbf{J}(\alpha)\{\mathbf{J}(\alpha)\}^T| \\
&\leq |\hat{\mathbf{J}}(\alpha)||\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)| + |\mathbf{J}(\alpha)||\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)| \\
&\leq c'|\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)| \ ,
\end{aligned}$$

where c' is some positive constant. The last inequality follows from the fact that the rows of $\mathbf{J}(\alpha)$ and $\hat{\mathbf{J}}(\alpha)$ are of unit length. The result now follows from Lemma 2.2.7. □

Lemma 2.2.9 *For $M > 0$, define $K_M^n = \{\alpha : \alpha \in S_n$, and $v(\alpha) \leq M\}$. Then $\sup_{\alpha \in K_M^n} |\hat{v}(\alpha) - v(\alpha)|$ converges in probability to zero as n tends to infinity.*

Proof : From Lemma 2.2.8, it is easy to see that

$$\sup_{\alpha \in S_n} |\hat{\mathbf{D}}(\alpha) - \mathbf{D}(\alpha)| \xrightarrow{p} 0 \,,$$

$$\sup_{\alpha \in S_n} |[\det\{\hat{\mathbf{J}}(\alpha)\}]^2 - [\det\{\mathbf{J}(\alpha)\}]^2| \xrightarrow{p} 0$$

and

$$\sup_{\alpha \in S_n} |\det\{\hat{\mathbf{D}}(\alpha)\} - \det\{\mathbf{D}(\alpha)\}| \xrightarrow{p} 0 \quad as\ n \longrightarrow \infty \,.$$

Next, note that there exists $\delta > 0$ such that, for any $\alpha \in K_M^n$, $[\det\{\mathbf{J}(\alpha)\}]^2 > \delta$. The existence of such a δ follows from some routine analysis using some of the arguments in the proof of Theorem 2.2.3. So, for sufficiently large n, with probability tending to one we have, $[\det\{\hat{\mathbf{J}}(\alpha)\}]^2 > \delta$. Therefore, for $\alpha \in K_M^n$,

$$\begin{aligned} |\hat{v}(\alpha) - v(\alpha)| &\leq \frac{|\det\{\hat{\mathbf{D}}(\alpha)\} - \det\{\mathbf{D}(\alpha)\}|}{[\det\{\hat{\mathbf{J}}(\alpha)\}]^2} \\ &+ \frac{|\det\{\mathbf{D}(\alpha)\}|\,|[\det\{\hat{\mathbf{J}}(\alpha)\}]^2 - [\det\{\mathbf{J}(\alpha)\}]^2|}{[\det\{\hat{\mathbf{J}}(\alpha)\}]^2[\det\{\mathbf{J}(\alpha)\}]^2} \\ &\leq \frac{|\det\{\hat{\mathbf{D}}(\alpha)\} - \det\{\mathbf{D}(\alpha)\}| + |[\det\{\hat{\mathbf{J}}(\alpha)\}]^2 - [\det\{\mathbf{J}(\alpha)\}]^2|}{\delta^2} \,. \end{aligned}$$

Hence, we have the result. □

Proof of Theorem 2.2.6 : From Theorem 2.2.5, we have that the α^*, which minimizes $v(\alpha)$, is in the set K_M^n, and hence in view of Lemma 2.2.9 $\hat{\alpha}$ will be in K_M^n with probability tending to one as n tends to infinity if $M > 0$ is chosen to be suitably large.

Next, since $\hat{\alpha}$ minimizes $\hat{v}(\alpha)$, and α^* minimizes $v(\alpha)$, it follows by some straightforward analysis that $|\hat{v}(\hat{\alpha}) - v(\hat{\alpha})| < \epsilon$ and $|\hat{v}(\alpha^*) - v(\alpha^*)| < \epsilon$ will imply that $|\hat{v}(\hat{\alpha}) - v(\alpha^*)| < \epsilon$. Hence

$$Pr\{|\hat{v}(\hat{\alpha}) - v(\alpha^*)| > \epsilon\} \leq Pr\{|\hat{v}(\hat{\alpha}) - v(\hat{\alpha})| > \epsilon\} + Pr\{|\hat{v}(\alpha^*) - v(\alpha^*)| > \epsilon\}$$

At this point, it follows from Lemma 2.2.9 that $\hat{v}(\hat{\alpha}) - v(\alpha^*)$ converges *in probability* to zero. The proof is now complete after observing the inequality

$$|v(\hat{\alpha}) - v(\alpha^*)| \leq |v(\hat{\alpha}) - \hat{v}(\hat{\alpha})| + |\hat{v}(\hat{\alpha}) - v(\alpha^*)|$$

and using Lemma 2.2.9. □

It follows from the above two Theorems that both of $v(\alpha^*)$ and $v(\hat{\alpha})$ converge to one, which is the lower bound discussed in the previous section following Theorem 2.2.3. Recall from this discussion that the asymptotic generalized variance of $\hat{\theta}_M^{(\alpha)}$ is $(c/n)^d \det(\Sigma) v(\alpha)$. Consequently, it now follows from Theorems 2.2.5 and 2.2.6 that the adaptive selection of $\alpha \in S_n$ will produce an estimate with asymptotic generalized variance $(c/n)^d \det(\Sigma)$. As noted by Bickel (1964) and Babu and Rao (1988), the asymptotic generalized variance of the vector of coordinatewise medians is $(c/n)^d \det(\Gamma)$, where the (i,j)-th element of Γ is $(\sigma_{ii}\sigma_{jj})^{1/2}(2/\pi)\sin^{-1}\rho_{ij}$, $\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2}$, σ_{ij} is the (i,j)-th element of Σ and c is as defined in Theorem 2.2.2. Following the line of arguments used in the proof of Theorem 2.2.3 it is easy to see that $\det(\Gamma) \geq \det(\Sigma)$, and equality holds only if Σ is a diagonal matrix. If the asymptotic efficiency of two competing estimates of a d-dimensional location parameter is now defined as the d-th root of the ratio of their asymptotic generalized variances, the efficiency of our adaptive equivariant estimate compared to the nonequivariant vector of medians is always greater than or equal to one. Further, the asymptotic efficiency of our estimate compared to the usual vector of means is the same as the efficiency of sample median compared to sample mean in the univariate problem, and it may be greater or smaller than one depending on the nature of the tail of the univariate marginal g of the d-variate spherically symmetric density f. These critical observations enable us to get a good feeling of the subtle and intriguing connection between affine equivariance and asymptotic efficiency of multivariate versions of median when there exist correlations among the observed variables.

We close this section by presenting some simulation results to demonstrate the performance of the adaptive equivariant estimate in small samples. We have generated observations from bivariate normal [i.e. $h(x,y) = (2\pi)^{-1}\exp(-(x^2+y^2)/2)$] and Laplace [i.e. $h(x,y) = (2\pi)^{-1}\exp(-\sqrt{x^2+y^2})$] distributions with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $\theta = (0,0)^T$. We have used a set of five different values of ρ and two sample sizes, namely 20 and 30. Our adaptive equivariant estimate was compared with the nonequivariant vector of medians, and for the purpose of efficiency computation, the estimates of their generalized variances were computed based on 2000 Monte Carlo replications. The

efficiency is taken to be the square root of the ratio of the generalized variances of the two competing bivariate location estimates.

Table 2.1: Efficiency figures for bivariate normal

Sample Size	ρ 0.75	0.80	0.85	0.90	0.95
20	1.1039	1.1876	1.2657	1.3702	1.6202
30	1.1447	1.2637	1.3031	1.3882	1.6849

Table 2.2: Efficiency figures for bivariate Laplace

Sample Size	ρ 0.75	0.80	0.85	0.90	0.95
20	1.0679	1.1035	1.1611	1.2533	1.4819
30	1.0746	1.1659	1.2314	1.4326	1.7864

It is apparent from Tables 2.1 and 2.2 that even with small sample sizes, there is a gain in efficiency when the adaptive equivariant estimate is used instead of the nonequivariant vector of univariate medians if the correlation between the variables is high. As ρ increases efficiency increases, and there is an increase in efficiency with increase in the sample size too. We would like to point out that in small samples, the gain in efficiency for the adaptive equivariant estimate seems to be more in the bivariate normal case than in the bivariate Laplace case.

2.2.3 Some Real Examples

In this section, we will consider two real data sets and explore the impact of adaptive transformation and retransformation strategy on their analysis. In both the examples we will estimate the generalized variances of the location estimates by the bootstrap method [see e.g. Efron (1982)]. One of the primary motivations behind considering the transformation retransformation estimate is that once we have the desired transformation matrix $\mathbf{X}(\alpha)$, it is quite easy to compute the estimate as it involves only determining the vector of coordinatewise medians of the transformed observations $\{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_i$'s and then retransforming that vector of univariate medians. As a consequence, one can conveniently estimate the conditional generalized variance of the transformation retransformation estimate using the bootstrap method once $\alpha \in S_n$ is fixed and the transformation matrix is formed. In each

case considered here, we have used 10,000 bootstrap replications to estimate the generalized variance, and it took only a negligible amount of time on a 486 PC equipped with a standard FORTRAN compiler. We would like to note here that the sampling variation of any other affine equivariant multivariate median proposed in the literature [e.g. Tukey (1975), Oja (1983) and Liu (1990)] is extremely difficult to estimate from the data. It is virtually impossible to use the bootstrap or other resampling techniques for any of them in practice due to the complex computational problems associated with each of them in the case of high or even moderately high dimensional data.

Example 2.2 : This example deals with the famous Iris data analyzed by R. A. Fisher and many eminent statisticians assuming multivariate normality. We have applied our technique of adaptive transformation and retransformation to all three different species considered in this data set, namely *Iris Setosa*, *Iris Versicolour* and *Iris Virginica*. Each data point in the set is four dimensional with variables : sepal length, sepal width, petal length and petal width, and there are 50 observations for each species. Table 2.3 gives the adaptive transformation retransformation medians and their estimated root mean squared errors (RMSE) for these variables separately for three different species.

Table 2.3: Transformation retransformation medians and their estimated RMSE's for Iris data

Species	sepal length	sepal width	petal length	petal width
Setosa	4.99	3.39	1.46	0.23
	(0.0690)	(0.0704)	(0.0285)	(0.0161)
Virginica	6.4456	2.9658	5.4039	2.0434
	(0.1264)	(0.0534)	(0.0769)	(0.0640)
Versicolour	6.0355	2.8285	4.3511	1.3482
	(0.1319)	(0.0549)	(0.0973)	(0.0475)

The estimated correlation matrices of the sample medians for three Iris species are :

$$\begin{pmatrix} 1.0 & 0.81 & 0.33 & 0.25 \\ & 1.0 & 0.22 & 0.27 \\ & & 1.0 & 0.31 \\ & & & 1.0 \end{pmatrix} \quad \begin{pmatrix} 1.0 & 0.50 & 0.75 & 0.24 \\ & 1.0 & 0.61 & 0.72 \\ & & 1.0 & 0.53 \\ & & & 1.0 \end{pmatrix} \quad \begin{pmatrix} 1.0 & 0.78 & 0.72 & 0.52 \\ & 1.0 & 0.79 & 0.74 \\ & & 1.0 & 0.84 \\ & & & 1.0 \end{pmatrix}$$

In addition to the adaptive equivariant estimate, we have computed the nonequivariant vector of medians and estimated the generalized variances for both of them in each species in order to make a comparison. Interestingly, the equivariant estimate turns out to be more efficient than the non-equivariant one for *Iris Versicolour* and *Iris Virginica* (estimated

efficiencies being 1.9158 and 1.8259 respectively in the two cases), while it turns out to be less efficient in the case of *Iris Setosa* (estimated efficiency being only 0.8522).

Example 2.3 : The data set used in this example was originally obtained from the laboratory of James S. Elliot, M.D. of the Urology Section, Veteran's Administration Medical Center, Palo Alto, California and the Division of Urology, Stanford University School of Medicine, Stanford, California, and it is reported in Andrews and Herzberg (1985). We have considered four physical characteristics of 33 urine specimens with calcium oxalate crystals (see Table 2.4). These variables are : specific gravity (i.e. the density of urine relative to water), pH (i.e. the negative logarithm of the hydrogen ion concentration), osmolarity (which is proportional to the concentration of molecules in the solution) and conductivity (which is proportional to the concentration of charged ion in the solution). As one would expect, the correlations among these variables are fairly high and the estimated efficiency of the adaptive equivariant estimate compared to the nonequivariant vector of medians turns out to be 2.2870. In other words, the transformation retransformation strategy significantly reduces the sampling variation in the location estimate in this case. The transformation and retransformation medians and their estimated root mean squared errors and correlation matrix are presented in Table 2.5.

It is clear from the preceding two examples that one does sometimes (though not always) gain by using the adaptive equivariant estimate. Our analysis enables us to choose between the equivariant transformation retransformation median and the nonequivariant vector of usual medians using a simple and convenient rule after the sampling variations of the two multivariate location estimates are estimated from the data.

2.3 Spatial Median

Let us now consider the spatial median $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ of the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s by minimizing the sum $\sum_{j\not\in\alpha} \|\boldsymbol{Y}_j^{(\alpha)} - \boldsymbol{\phi}\|$, where for $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_d^2}$. Finally, in order to express things back in terms of the original coordinate system, we need to retransform $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ into $\hat{\boldsymbol{\theta}}_S^{(\alpha)} = \mathbf{X}(\alpha)\hat{\boldsymbol{\phi}}_n^{(\alpha)}$, which is our desired location estimate. In view of its construction, it is clear that $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$ is an affine equivariant estimate of location, but we need to settle the question of how to choose the 'data driven coordinate system' or equivalently the data based transformation matrix $\mathbf{X}(\alpha)$. An answer to this question is provided in the following Theorem.

Let us write $h(\boldsymbol{x})$ to denote the elliptically symmetric density $\{\det(\Sigma)\}^{-1/2} f(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x})$, where Σ is a $d\times d$ positive definite matrix and $f(\boldsymbol{x}^T\boldsymbol{x})$ is a continuous spherically symmetric density around the origin in $\mathbb{R}^d$. The $\boldsymbol{X}_i$'s will be assumed to be i.i.d observations with

Table 2.4: Physical characteristics of urine specimens with calcium oxalate crystals

Specific gravity	pH	Osmolarity	Conductivity
1.021	5.94	774	27.9
1.024	5.77	698	19.5
1.024	5.60	866	29.5
1.021	5.53	775	31.2
1.024	5.36	853	27.6
1.026	5.16	822	26.0
1.013	5.86	531	21.4
1.010	6.27	371	11.2
1.011	7.01	443	21.4
1.011	6.13	364	10.9
1.031	5.73	874	17.4
1.020	7.94	567	19.7
1.040	6.28	838	14.3
1.021	5.56	658	23.6
1.025	5.71	854	27.0
1.026	6.19	956	27.6
1.034	5.24	1236	27.3
1.033	5.58	1032	29.1
1.015	5.98	487	14.8
1.013	5.58	516	20.8
1.014	5.90	456	17.8
1.012	6.75	251	5.1
1.025	6.90	945	33.6
1.026	6.29	833	22.2
1.028	4.76	312	12.4
1.027	5.40	840	24.5
1.018	5.14	703	29.0
1.022	5.09	736	19.8
1.025	7.90	721	23.6
1.017	4.81	410	13.3
1.024	5.40	803	21.8
1.016	6.81	594	21.4
1.015	6.03	416	12.8

Table 2.5: Transformation retransformation medians, their estimated RMSE's and correlations for urine data

Variables	Median	Correlation matrix			
Specific gravity	1.0222 (0.0015)	1.00	-0.1161	0.9207	0.5223
pH	5.8718 (0.1253)		1.00	-0.2217	-0.4135
Osmolarity	730.1650 (55.3338)			1.00	0.7599
Conductivity	21.6264 (1.7926)				1.00

common elliptically symmetric density $h(\boldsymbol{x}-\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the location of elliptic symmetry for the data.

Theorem 2.3.1 *For any given subset α of $\{1,\ldots,n\}$ with size $d+1$ and given the $\mathbf{X}_i$'s with $i \in \alpha$, the conditional asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_S^{(\alpha)}-\boldsymbol{\theta})$ is d-variate normal with zero mean and a variance covariance matrix $\Delta\{f,\Sigma,\mathbf{X}(\alpha)\}$ that depends on f, Σ and the transformation matrix $\mathbf{X}(\alpha)$. Here the positive definite matrix Δ is such that the difference $\Delta\{f,\Sigma,\mathbf{A}\}-\Delta\{f,\Sigma,\mathbf{B}\}$ is non-negative definite (i.e. $\Delta\{f,\Sigma,\mathbf{A}\} \geq_{n.n.d} \Delta\{f,\Sigma,\mathbf{B}\}$) for any f, Σ and any two $d\times d$ invertible matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{B}^T\Sigma^{-1}\mathbf{B}=\lambda\mathbf{I}_d$, where $\lambda>0$ is a constant and $\mathbf{I}_d$ is the $d\times d$ identity matrix. Further, for any such $\mathbf{B}$, we have $\Delta\{f,\Sigma,\mathbf{B}\}=c(d,f)\Sigma$, where $c(d,f)=\pi^{-1}d(d-1)^{-2}\{g(0)\}^{-2}[\Gamma\{(d-1)/2\}]^{-2}\{\Gamma(d/2)\}^2$, g being the univariate marginal of the spherically symmetric density f on $\mathbb{R}^d$.*

Proof: First observe that in view of affine equivariance of $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$, it is enough to consider the case when $\boldsymbol{\theta}=\mathbf{0}$ and $\Sigma=\mathbf{I}_d$. Then $h(\boldsymbol{x}-\boldsymbol{\theta})$ reduces to the spherically symmetric density $f(\boldsymbol{x}^T\boldsymbol{x})$. Now, for a given subset α with size $d+1$ of $\{1,\ldots,n\}$ and given the $\boldsymbol{X}_i$'s for which $i \in \alpha$, the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s are conditionally independent, and they are identically distributed with common elliptically symmetric density $h\{y|\mathbf{X}(\alpha)\}=|\det\{\mathbf{X}(\alpha)\}|f[y^T\{\mathbf{X}(\alpha)\}^T\mathbf{X}(\alpha)y]$. For a random vector $\boldsymbol{Y}$ with density h, elliptic symmetry around the origin implies that the distribution of $\|\boldsymbol{Y}\|^{-1}\boldsymbol{Y}$ does not depend on f but on $\mathbf{X}(\alpha)$. Consider the matrices $\mathbf{C}\{\mathbf{X}(\alpha)\}=E_h(\|\boldsymbol{Y}\|^{-2}\boldsymbol{Y}\boldsymbol{Y}^T)$ and $\mathbf{D}\{f,\mathbf{X}(\alpha)\}=E_h\{\|\boldsymbol{Z}\|^{-1}(\mathbf{I}_d-\|\boldsymbol{Y}\|^{-2}\boldsymbol{Y}\boldsymbol{Y}^T)\}$. Then it follows from Chaudhuri (1992a) that given $\mathbf{X}(\alpha)$, the conditional limiting distribution of $n^{1/2}\hat{\boldsymbol{\phi}}_n^{(\alpha)}$, where $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ is the spatial median based on the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s, is normal with zero mean and $[\mathbf{D}\{f,\mathbf{X}(\alpha)\}]^{-1}\mathbf{C}\{\mathbf{X}(\alpha)\}[\mathbf{D}\{f,\mathbf{X}(\alpha)\}]^{-1}$ as the variance covariance matrix. Further, elliptic symmetry of h around the origin implies that $\mathbf{D}\{f,\mathbf{X}(\alpha)\}=\mu(d,f)\mathbf{G}\{\mathbf{X}(\alpha)\}$, where μ

is a positive constant depending on dimension d and f, and $\mathbf{G}$ is a positive definite symmetric matrix depending on $\mathbf{X}(\alpha)$ only. Finally, since $\hat{\theta}_S^{(\alpha)} = \mathbf{X}(\alpha)\hat{\phi}_n^{(\alpha)}$, the conditional limiting distribution of $n^{1/2}\hat{\theta}_S^{(\alpha)}$ must be normal with variance covariance matrix

$$\mathbf{X}(\alpha)[\mathrm{D}\{f, \mathbf{X}(\alpha)\}]^{-1}\mathrm{C}\{\mathbf{X}(\alpha)\}[\mathrm{D}\{f, \mathbf{X}(\alpha)\}]^{-1}\{\mathbf{X}(\alpha)\}^T$$

$$= \{\mu(d, f)\}^{-2}\mathbf{X}(\alpha)[\mathbf{G}\{\mathbf{X}(\alpha)\}]^{-1}\mathrm{C}\{\mathbf{X}(\alpha)\}[\mathbf{G}\{\mathbf{X}(\alpha)\}]^{-1}\{\mathbf{X}(\alpha)\}^T,$$

which we can write as $\Delta\{f, \mathrm{I}_d, \mathbf{X}(\alpha)\}$, where by affine equivariance we have $\Delta\{f, \Sigma, \mathbf{A}\} = \Sigma^{1/2}\Delta\{f, \mathrm{I}_d, \Sigma^{-1/2}\mathbf{A}\}\Sigma^{1/2}$.

Next observe that it is enough to prove the non-negative definite ordering of Δ stated in the Theorem for $\Sigma = \mathrm{I}_d$ and $\mathbf{B}^T\mathbf{B} = \mathrm{I}_d$ because when $\mathbf{B}^T\mathbf{B} = \lambda\mathrm{I}_d$, $\Delta\{f, \mathrm{I}_d, \mathbf{B}\}$ is a diagonal matrix that does not depend on the value of λ or the specific choice of $\mathbf{B}$. Also, for any nonsingular $\mathbf{A}$,

$$\Delta\{f, \mathrm{I}_d, \mathbf{A}\} = \{\mu(d, f)\}^{-2}\mathbf{A}\{\mathbf{G}(\mathbf{A})\}^{-1}\mathrm{C}(\mathbf{A})\{\mathbf{G}(\mathbf{A})\}^{-1}\mathbf{A}^T,$$

and hence in order to prove the non-negative definite ordering of Δ, we can choose f to be any specific density as its effect appears only through the scalar factor $\mu(d, f)$. In particular, we can choose $f(\boldsymbol{x}^T\boldsymbol{x})$ to be the multivariate Laplace density $k\exp\{-(\boldsymbol{x}^T\boldsymbol{x})^{1/2}\}$. Then it is straight forward to verify that for a random vector $\boldsymbol{Y}$ with density $h(\boldsymbol{y}|\mathbf{A}) = \det(\mathbf{A})f(\boldsymbol{y}^T\mathbf{A}^T\mathbf{A}\boldsymbol{y})$, we must have

$$(\mathbf{A}^T)^{-1}\mathrm{D}(f, \mathbf{A})\mathbf{A}^{-1} = (\mathbf{A}^T)^{-1}E\{\|\boldsymbol{Y}\|^{-1}(\mathrm{I}_d - \|\boldsymbol{Y}\|^{-2}\boldsymbol{Y}\boldsymbol{Y}^T)\}\mathbf{A}^{-1}$$

$$= \mathbf{COV}\left\{\|\boldsymbol{Y}\|^{-1}(\mathbf{A}^T)^{-1}\boldsymbol{Y}\ ,\ \|\mathbf{A}\boldsymbol{Y}\|^{-1}\mathbf{A}\boldsymbol{Y}\right\},$$

where $\mathbf{COV}$ denotes the covariance matrix between two random vectors. Also, note that $(\mathbf{A}^T)^{-1}\mathrm{C}(\mathbf{A})\mathbf{A}^{-1}$ is nothing but the dispersion matrix of $\|\boldsymbol{Y}\|(\mathbf{A}^T)^{-1}\boldsymbol{Y}$. Now, the non-negative definiteness of the difference

$$\Delta(f, \mathrm{I}_d, \mathbf{A}) - \Delta(f, \mathrm{I}_d, \mathbf{B}) =$$

$$\mathbf{A}\{\mathrm{D}(f, \mathbf{A})\}^{-1}\mathrm{C}(\mathbf{A})\{\mathrm{D}(f, \mathbf{A})\}^{-1}\mathbf{A}^T - \mathbf{B}\{\mathrm{D}(f, \mathbf{B})\}^{-1}\mathrm{C}(\mathbf{B})\{\mathrm{D}(f, \mathbf{B})\}^{-1}\mathbf{B}^T$$

follows from the simple fact that for any two d-dimensional random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$, the difference

$$\{\mathbf{COV}(\boldsymbol{V}, \boldsymbol{U})\}^{-1}\mathbf{DISP}(\boldsymbol{V})\{\mathbf{COV}(\boldsymbol{U}, \boldsymbol{V})\}^{-1} - \{\mathbf{DISP}(\boldsymbol{U})\}^{-1}$$

is non-negative definite, where **DISP** denotes dispersion matrix, and all the matrices involved are invertible. Finally, the expression of $c(d, f)$ stated in the Theorem follows

from a direct algebraic computation using the asymptotic distribution of spatial median in spherically symmetric models [see e.g. Brown (1983), Chaudhuri (1992a)]. □

The main message communicated by the above Theorem is that we need to choose $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a matrix of the form $\lambda\mathbf{I}_d$, which is a diagonal matrix with all diagonal entries equal. In other words, the coordinate system represented by the transformation matrix $\Sigma^{-1/2}\mathbf{X}(\alpha)$ should be as orthonormal in nature as possible. The expression for $c(d, f)$ in Theorem 2.3.1 implies that when $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ is chosen to be close to a diagonal matrix with all diagonal entries equal, the asymptotic efficiency of the estimate $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$ becomes close to that of the spatial median under spherically symmetric models (i.e. when $\Sigma = \sigma^2\mathbf{I}_d$), and it will be more efficient than spatial median in elliptically symmetric models [see Chaudhuri (1992a)]. It is known that for spherically symmetric data rotationally equivariant spatial median is more efficient than the vector of coordinatewise medians, which lacks rotational equivariance [see Brown (1983), Chaudhuri (1992a)], and with a proper selection of $\mathbf{X}(\alpha)$, $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$ too will have similar superior performance. Another implication of Theorem 2.3.1 is that with appropriate choice of $\mathbf{X}(\alpha)$, the estimate $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$ will be more (or less) efficient than the sample mean vector depending on whether the tail of the density f is 'heavy' (or 'light').

We will discuss a procedure for choosing the transformation matrix $\mathbf{X}(\alpha)$ from the data when we will discuss numerical examples in the next subsection, and there we will compare the finite sample performance of our estimate with that of some other well-known estimates of multivariate location. But before that let us close this section by noting that an alternative affine equivariant modification of spatial median has been considered in the literature by other authors [see e.g. Isogai (1985), Rao (1988)], who computed spatial median based on multivariate observations transformed by the square root of the usual variance covariance matrix. While transforming the data points by the square root of the sample variance covariance matrix is a popular approach, the resulting coordinate system does not have any simple and natural geometric interpretation as we have already pointed out. We will see in the next chapter that the strategy of transforming the data points using an appropriately chosen $\mathbf{X}(\alpha)$ leads to an affine invariant modification of the well-known angle test, which turns out to be 'distribution free' in nature in the sense that the null distribution of the test statistic under elliptically symmetric model does not depend on the unknown density f. This is not achievable by transforming the observations using the square root of the sample variance covariance matrix.

2.3.1 Simulation Studies and Data Analysis

It is quite clear from our main results and discussion in the preceding subsection that we need to choose $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a matrix of the form $\lambda\mathbf{I}_d$. Since Σ will be unknown in practice, we have to estimate that from the data and we will need a consistent and affine equivariant estimate (say $\hat{\Sigma}$). When the variables observed in the data have finite population variances, we can use the usual variance covariance matrix for this purpose. In any case, after obtaining $\hat{\Sigma}$, we will try to choose $\mathbf{X}(\alpha)$ in such a way that the eigenvalues of the positive definite matrix $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ become as equal as possible. To achieve this, our strategy will be to minimize either the ratio between the arithmetic mean and the geometric mean or that between the geometric mean and the harmonic mean of the eigenvalues. Note that a major advantage in using such a criterion is that it does not involve explicit computation of the eigenvalues of the matrix. Arithmetic and harmonic means of the eigenvalues can be obtained from the trace of the matrix and that of its inverse respectively, while the geometric mean can be computed from its determinant. In our numerical studies, we have observed that the criteria based on different ratios yield more or less similar results. Instead of minimizing the ratio over all possible subsets α with size $d+1$ of $\{1,\ldots,n\}$, one can substantially reduce the amount of computation by stopping the search for optimal $\mathbf{X}(\alpha)$ as soon as the ratio becomes smaller than $1+\epsilon$, where ϵ is a preassigned small positive number. In our simulations and data analysis, we did not observe such an approach to cause any significant change in the statistical performance of the procedures though there was considerable gain in the speed of computation. Of course, there are other different ways to achieve this goal of making $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ as close as possible to a diagonal matrix with all diagonal entries equal. We have adopted a specific strategy that is computationally convenient and has been observed to work fairly well in our numerical investigations. Note that once $\mathbf{X}(\alpha)$ is chosen, we can compute the spatial median $\hat{\Phi}_n^{(\alpha)}$ from the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s using any of the standard algorithms discussed in the literature [see e.g. Gower (1974), Chaudhuri (1996)].

We will now discuss a simulation study that was undertaken with the objective of comparing the finite sample performance of $\hat{\boldsymbol{\theta}}_S^{(\alpha)}$ with that of the sample mean vector and the vector of coordinatewise sample medians. We have used sample size $n = 30$ and considered the cases $d = 2$ and 3. We generated data from three different distributions, namely multivariate normal, multivariate Laplace (i.e. when $f(\boldsymbol{x}^T\boldsymbol{x}) = k\exp\{-(\boldsymbol{x}^T\boldsymbol{x})^{1/2}\}$) and multivariate t with 3 degrees of freedom. Keeping in mind location equivariance as well as equivariance under coordinatewise scale transformation of each of the three multivariate location estimates considered, we decided to generate data from the elliptically symmetric

density $h(x-\theta)$, where θ was taken to be the zero vector and Σ was taken to be the matrix with each diagonal entry equal to one and each off-diagonal entry equal to ρ. The value of ρ was chosen from the interval $[0,1)$. We will denote by e_1 and e_2 the efficiencies of $\hat{\theta}_S^{(\alpha)}$ compared with sample mean vector and the vector of coordinatewise sample medians respectively. For two competing estimates $\hat{\phi}_1$ and $\hat{\phi}_2$ of a d-dimensional location parameter ϕ, we will define the efficiency of the former estimate compared with the latter one as the d-th root of the ratio between $\det\{E(\hat{\phi}_2-\phi)(\hat{\phi}_2-\phi)^T\}$ and $\det\{E(\hat{\phi}_1-\phi)(\hat{\phi}_2-\phi)^T\}$ [see e.g. Bickel (1964)]. The results are reported in Tables 2.6 and 2.7. In each case, we have estimated the efficiencies e_1 and e_2 based on 10,000 Monte Carlo replications for $d=2$ and using 5,000 Monte Carlo replications for $d=3$. Since both of $\hat{\theta}_S^{(\alpha)}$ and sample mean vector are affine equivariant estimates, the value of e_1 remains constant for different values of ρ. The superior performance of $\hat{\theta}_S^{(\alpha)}$ for non-normal elliptically symmetric distributions (especially when ρ is large) is quite transparent in the figures given in Tables 2.6 and 2.7.

Table 2.6: Finite sample efficiency of affine equivariant modification of spatial median for $n=30$ and $d=2$

Distribution		ρ					
		0.00	0.75	0.80	0.85	0.90	0.95
Normal	e_1	0.7153	0.7153	0.7153	0.7153	0.7153	0.7153
	e_2	1.1313	1.4418	1.5243	1.6447	1.8285	2.1747
Laplace	e_1	1.2849	1.2849	1.2849	1.2849	1.2849	1.2849
	e_2	1.0861	1.3779	1.4655	1.5877	1.7688	2.1172
t with 3 d.f.	e_1	1.7676	1.7676	1.7676	1.7676	1.7676	1.7676
	e_2	1.0628	1.3551	1.4379	1.5512	1.7291	2.0769

Table 2.7: Finite sample efficiency of affine equivariant modification of spatial median for $n=30$ and $d=3$.

Distribution		ρ					
		0.00	0.75	0.80	0.85	0.90	0.95
Normal	e_1	0.7319	0.7319	0.7319	0.7319	0.7319	0.7319
	e_2	1.1649	1.5883	1.7140	1.8725	2.1219	2.6873
Laplace	e_1	1.1023	1.1023	1.1023	1.1023	1.1023	1.1023
	e_2	1.1701	1.6078	1.7271	1.9041	2.1757	2.7461
t with 3 d.f.	e_1	1.6725	1.6725	1.6725	1.6725	1.6725	1.6725
	e_2	1.1395	1.5725	1.6830	1.8538	2.1097	2.6413

Let us next consider two real data sets and try to investigate the performance of $\hat{\theta}_S^{(\alpha)}$ there. One of the primary reasons for using the transformation retransformation

technique is that once the optimal data based transformation matrix $\mathbf{X}(\alpha)$ is chosen, it is quite easy to compute $\hat{\theta}_S^{(\alpha)}$ as it requires only the computation of spatial median based on the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s. An important consequence of this is that one can conveniently use resampling techniques such as the bootstrap [see e.g. Efron (1982)] to estimate the conditional sampling variation of $\hat{\theta}_S^{(\alpha)}$ given the $\boldsymbol{X}_i$'s with $i \in \alpha$ (i.e. after $\mathbf{X}(\alpha)$ is fixed). In each of the two examples discussed below, we have used 10,000 bootstrap replications to estimate the sampling variation and the efficiency of our transformation retransformation estimate, and it took only a few seconds on a workstation equipped with a standard FORTRAN compiler.

Table 2.8: Transformation retransformation estimates and the results of bootstrap analysis of Fisher's Iris data

Species	Estimates and estimated RMSE's				Estimated efficiency
	Sepal length	Sepal width	Petal length	Petal width	
Setosa	5.0148	3.4180	1.4684	0.2376	$e_1^* = 1.0308$
	(0.0488)	(0.0648)	(0.0221)	(0.0137)	$e_2^* = 1.2482$
Versicolor	5.9111	2.8001	4.2733	1.3256	$e_1^* = 0.6607$
	(0.1178)	(0.0656)	(0.0961)	(0.0422)	$e_2^* = 2.4361$
Virginica	6.5421	2.9864	5.4953	2.0428	$e_1^* = 0.7494$
	(0.0926)	(0.0516)	(0.0802)	(0.0514)	$e_2^* = 1.9220$

Example 2.4 : Like Example 2.2 this example deals with the Iris data. Table 2.8 gives our location estimate and its root mean squared error (RMSE) as estimated by the bootstrap for each variable separately for different species. We have denoted by e_1^* and e_2^* the bootstrap estimates of the efficiencies of our affine equivariant modification of spatial median as compared with the sample mean vector and the vector of coordinatewise sample medians respectively. It is interesting to note that while there is a gain in efficiency when compared with the non-equivariant vector of coordinatewise median in all three species, when compared with the affine equivariant sample mean, there is gain only in the case of Iris Setosa, and there is a definite loss in efficiency in each of the other two cases. The entire analysis seems to make a very good case for using affine equivariant procedures.

Example 2.5 : The data set used in this example is the same as that in Example 2.3. Table 2.9 summarizes the results of the bootstrap analysis of this data set. The values of e_1^* and e_2^* indicate considerable gain in efficiency over the non-equivariant vector of coordinatewise medians and a small gain over the sample mean vector.

Table 2.9: Transformation retransformation estimates and the results of bootstrap analysis of urine data

Estimates and estimated RMSE's				Estimated
Specific Gravity	pH	Conductivity	Osmolarity	efficiency
1.025	7.90	721.0	23.6	$e_1^* = 1.0921$
(0.0016)	(0.1201)	(54.5453)	(1.6232)	$e_2^* = 2.8250$

2.4 Hodges–Lehmann Type Estimates

As in previous sections, define the transformation matrix $\mathbf{X}(\alpha)$ and transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s, for $1 \leq j \leq n$, $j \notin \alpha$. One can then compute coordinatewise HL-estimate $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ based on $\boldsymbol{Y}_j^{(\alpha)}$'s by minimizing the sum $\sum_{i \notin \alpha} \sum_{\substack{j \notin \alpha \\ j \neq i}} \left| \frac{\boldsymbol{Y}_j^{(\alpha)} + \boldsymbol{Y}_i^{(\alpha)}}{2} - \Phi \right|$ (here for $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, $|\boldsymbol{x}| = |x_1| + \ldots + |x_d|$). Finally, in order to express the estimate back in terms of the original coordinate system, we need to retransform $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ into $\hat{\boldsymbol{\theta}}_H^{(\alpha)} = \mathbf{X}(\alpha)\hat{\boldsymbol{\phi}}_n^{(\alpha)}$, which is our desired location estimate. Alternatively, one may also use Euclidean norm based procedures described in Chaudhuri (1992a) for constructing $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$. However, we will not consider that here. In view of the construction, it is obvious that $\hat{\boldsymbol{\theta}}_H^{(\alpha)}$ is equivariant under any arbitrary affine transformation of the data vectors, whatever be the procedure used to estimate $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$.

Theorem 2.4.1 *If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are independent and identically distributed with a common elliptically symmetric density $\{\det(\Sigma)\}^{-1/2} f[(\boldsymbol{x}-\boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\theta})]$, the conditional asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_H^{(\alpha)} - \boldsymbol{\theta})$ given the $\boldsymbol{X}_i$'s with $i \in \alpha$, is d-variate normal with zero mean and a variance covariance matrix $\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\}$ that depends on f, Σ and the transformation matrix $\mathbf{X}(\alpha)$ as $n \longrightarrow \infty$. Further, if the underlying distribution of $\boldsymbol{X}_i$'s is d-dimensional normal, the positive definite matrix $\mathcal{V}$ is such that $\det[\mathcal{V}\{f, \Sigma, \mathbf{A}\}] \geq \det[\mathcal{V}\{f, \Sigma, \mathbf{B}\}]$ for any Σ and any two $d \times d$ invertible matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{B}^T \Sigma^{-1} \mathbf{B}$ is a diagonal matrix.*

Proof: Note that, $\hat{\boldsymbol{\phi}}_n^{(\alpha)}$ is the coordinatewise HL-estimate based on the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s, $j \notin \alpha$. Then by the results of Bickel (1964), $n^{1/2}(\hat{\boldsymbol{\phi}}_n^{(\alpha)} - \boldsymbol{\phi})$ is asymptotically conditionally normal given $\mathbf{X}(\alpha)$ with mean zero, where $\boldsymbol{\phi} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{\theta}$. Now $\hat{\boldsymbol{\theta}}_H^{(\alpha)}$ is defined as $\{\mathbf{X}(\alpha)\}\hat{\boldsymbol{\phi}}_n^{(\alpha)}$. Thus the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_H^{(\alpha)} - \boldsymbol{\theta})$ given $\mathbf{X}(\alpha)$ is d-dimensional normal with mean zero. Straightforward algebra shows that the asymptotic variance covariance matrix to be $\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\}$ which depends on f, Σ and the transformation matrix $\mathbf{X}(\alpha)$. In fact, we do not need the assumption of elliptic symmetry

for asymptotic normality of our transformation retransformation HL-estimate. We made the assumption of elliptic symmetry for the asymptotic variance covariance matrix $\mathcal{V}$ to have the special structure

$$\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\} = \frac{1}{12}\{\int f_1^2(x)dx\}^{-2}\Sigma^{1/2}\{\mathrm{J}(\alpha)\}^{-1}\mathcal{U}(\alpha)[\{\mathrm{J}(\alpha)\}^T]^{-1}\Sigma^{1/2}, \tag{2.2}$$

where f_1 is the univariate marginal density of the spherically symmetric density f and the matrix $\mathrm{J}(\alpha)$ is obtained by normalizing the rows of $\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}^{-1}$ and $\mathcal{U}(\alpha)$ is the correlation matrix of $(F_1(Y_1), \ldots, F_d(Y_d))^T$. Here F_i is the i-th marginal distribution function of $\mathrm{J}(\alpha)\boldsymbol{X}$ and $\boldsymbol{Y}_i$ is the i-th coordinate variable of $\mathrm{J}(\alpha)\boldsymbol{X}$, where $\boldsymbol{X}$ has a spherically symmetric density f. The determinant ordering of the matrix $\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\}$ when the underlying distribution is d-dimensional normal follows from the arguments similar to those used in the proof of Theorem 2.2.3. □

In Section 2.3, it was suggested that one may select optimal α by choosing $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix with all diagonal entries equal. Following the arguments used in Theorem 2.2.6, it can be shown that for the above mentioned procedure, the conditional asymptotic generalized variance of the proposed estimate is $[12n\{\int f_1^2(x)dx\}^2]^{-d}\det\{\Sigma\}$. Thus the efficiency of the proposed estimate $\hat{\theta}_H^{(\alpha)}$ over sample mean is the same as that of the HL-estimate of location over mean in the univariate setup. Next we will discuss some simulation studies in an attempt to see the performance of the transformation retransformation HL-estimate $\hat{\theta}_H^{(\alpha)}$ in finite samples for different elliptically symmetric distributions.

2.4.1 Finite Sample Efficiency of Multivariate Hodges-Lehmann Estimate

We conclude this Section with a small simulation study on the efficiency of our affine equivariant HL-estimate of location. To determine the efficiency of a multivariate location estimate over another, we have once again used the notion of efficiency introduced by Bickel (1964) based on the d-th root of the ratio of the generalized variances of the competing estimates, where d is the dimension of the data. We carried out the simulations for three multivariate distributions, namely multivariate normal, multivariate Laplace and multivariate t with 3 d.f. with sample size $n = 30$ and 5000 Monte Carlo replications. The efficiency of the affine equivariant HL-type estimate is computed over the usual sample mean, the vector of coordinatewise median, affine equivariant transformation retransformation coordinatewise median and affine equivariant transformation retransformation spatial median, which are denoted by e_1, e_2, e_3 and e_4 respectively, for different values of ρ. The results are summarized in Tables 2.10 and 2.11.

Table 2.10: Finite sample efficiency of affine equivariant HL-estimate of location for $n = 30$ and $d = 2$

Distribution		ρ					
		0.00	0.75	0.80	0.85	0.90	0.95
Laplace	e_1	1.0919	1.0919	1.0919	1.0919	1.0919	1.0919
	e_2	0.9047	1.1356	1.2078	1.3086	1.4629	1.7579
	e_3	1.2854	1.2854	1.2854	1.2854	1.2854	1.2854
	e_4	0.8360	0.8360	0.8360	0.8360	0.8360	0.8360
Normal	e_1	0.8480	0.8480	0.8480	0.8480	0.8480	0.8480
	e_2	1.3534	1.7095	1.8224	1.9734	2.1929	2.6099
	e_3	1.7757	1.7757	1.7757	1.7757	1.7757	1.7757
	e_4	1.2328	1.2328	1.2328	1.2328	1.2328	1.2328
t with 3 d.f.	e_1	1.5883	1.5883	1.5883	1.5883	1.5883	1.5883
	e_2	0.9474	1.2261	1.2954	1.3992	1.5603	1.8810
	e_3	1.3971	1.3971	1.3971	1.3971	1.3971	1.3971
	e_4	0.8861	0.8861	0.8861	0.8861	0.8861	0.8861

2.5 Concluding Remarks

Remark 1: It is interesting to note that the procedure proposed here for selection of α and $\mathbf{X}(\alpha)$ does not require any knowledge of the form of the underlying density f. As has been observed in previous sections, there is a nice and intuitively appealing geometric interpretation for such an approach. The matrix $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes a diagonal matrix when the columns of $\Sigma^{-1/2}\mathbf{X}(\alpha)$ are orthogonal to one another. In other words, our recommendation amounts to transforming the observation vectors using a new 'data driven coordinate system' determined by the transformation matrix $\mathbf{X}(\alpha)$ such that the coordinate system is as orthogonal as possible in a d-dimensional vector space, where the inner product and orthogonality are defined based on the positive definite scatter matrix Σ of the probability distribution associated with the data vectors.

Remark 2: As we have discussed in detail in Section 2.2, the concern about poor efficiency of the nonequivariant vector of univariate medians raised by Bickel (1964) and Brown and Hettmansperger (1987) can be settled by using our adaptive transformation and retransformation strategy. Asymptotically our equivariant estimate outperforms the nonequivariant vector of medians as well as the affine equivariant vector of means in the presence of correlation among the variables if the underlying distribution is elliptically symmetric with univariate marginals having heavy tails. Our simulation results amply indicate a gain in the efficiency over the vector of coordinatewise median even in finite

Table 2.11: Finite sample efficiency of affine equivariant HL-estimate of location for $n = 30$ and $d = 3$

Distribution		ρ					
		0.00	0.75	0.80	0.85	0.90	0.95
Laplace	e_1	0.9926	0.9926	0.9926	0.9926	0.9926	0.9926
	e_2	1.0585	1.4729	1.5763	1.7326	1.9993	2.4915
	e_3	1.2425	1.2425	1.2425	1.2425	1.2425	1.2425
	e_4	0.8807	0.8807	0.8807	0.8807	0.8807	0.8807
Normal	e_1	0.8344	0.8344	0.8344	0.8344	0.8344	0.8344
	e_2	1.3382	1.8066	1.9327	2.1351	2.3848	2.9853
	e_3	1.5036	1.5036	1.5036	1.5036	1.5036	1.5036
	e_4	1.1299	1.1299	1.1299	1.1299	1.1299	1.1299
t with 3 d.f.	e_1	1.4273	1.4273	1.4273	1.4273	1.4273	1.4273
	e_2	0.9495	1.2946	1.3882	1.5395	1.7494	2.1956
	e_3	1.2362	1.2362	1.2362	1.2362	1.2362	1.2362
	e_4	0.8203	0.8203	0.8203	0.8203	0.8203	0.8203

sample situations for standard elliptically symmetric distributions when the correlations among the variables in the data are significant.

Remark 3: When the underlying distribution deviates significantly from being elliptically symmetric, instead of minimizing $v(\alpha)$, one can try to estimate the generalized variance of the transformation retransformation median for a fixed α using some resampling technique, and then minimize that estimated variance w.r.t $\alpha \in S_n$. However, such an approach will be computationally quite intensive, and we will not discuss it here.

Remark 4: Chaudhuri (1992a) proposed and studied a class of HL-type estimates in multidimension, which are equivariant under rotations (i.e. under orthogonal transformations) but they are not equivariant under arbitrary affine transformations of the data. We can employ our transformation retransformation strategy in conjunction with those estimates to define another related class of affine equivariant versions of HL-type estimates retaining their good efficiency properties.

Remark 5: It is also worth noting here that the estimation of finite sample variation of our proposed transformation retransformation estimate is quite simple. After selecting the optimal transformation matrix, one can use any resampling techniques (e.g. bootstrap) on the transformed observations to estimate the variance covariance matrix of the proposed estimate. This simplicity in computing the finite sample variation of the estimate is a major advantage for the transformation retransformation estimate when it comes to

practical applications.

Remark 6: We have seen earlier that none of the affine equivariant multivariate medians proposed in the existing literature possess good breakdown property (i.e. 50% breakdown) of univariate median, whereas nonequivariant medians like vector of coordinatewise medians and spatial median have breakdown 50%. After we fix the transformation, we need to compute only nonequivariant vector of coordinatewise medians or spatial median. Thus our transformation retransformation estimates can only break if the selected transformation matrix breaks. But our selection procedure guarantees that the determinant of $\{\mathbf{X}(\alpha)\}^T\hat{\Sigma}^{-1}\mathbf{X}(\alpha)$ always remains finite and bounded away from zero. We only need a robust and consistent estimate $\hat{\Sigma}$ of the scatter matrix Σ. It is easy to observe that the breakdown point of $\hat{\Sigma}$ provides a lower bound for the breakdown point of the affine equivariant transformation retransformation estimates. Thus if one uses high breakdown affine equivariant estimates of Σ (see Davies 1987), the asymptotic breakdown of transformation retransformation spatial median or coordinatewise median would be 50%.

Chapter 3

Multivariate Sign and Rank Tests

3.1 Introduction

The simplicity and widespread popularity of univariate sign and rank tests for one-sample and two-sample location problems have motivated numerous statisticians to explore several possibilities for their multivariate generalization. In the late 50's and the early 60's Bennett (1962), Bickel (1965), Blumen (1958) and Chatterjee (1966) developed some multivariate sign and rank based methods. More recent attempts in that direction have been made by Brown and Hettmansperger (1987,1989), Brown, Hettmansperger, Nyblom and Oja (1992), Liu (1992), Liu and Singh (1993), Oja and Nyblom (1989), Randles (1989), Hettmansperger, Möttönen and Oja (1996a,b) and others. Readers are referred to Hettmansperger, Nyblom and Oja (1992) and Chaudhuri and Sengupta (1993) for some recent detailed reviews. The popularity of univariate sign and rank based methods have their root in their distribution free nature and their applicability to solve a number of practical problems for which more traditional techniques cannot be used as they frequently require the assumption of normality of the data, which may be hard to justify in practice. It is a well-known fact that sample median is the estimate of location naturally associated with the univariate sign test, and in the same way Hodges-Lehmann (1963) estimate is naturally associated with Wilcoxon's signed rank test. So it is reasonable to expect that a multivariate version of median will be associated with a multivariate sign test, and a multivariate analog of Hodges-Lehmann (HL) estimate will be associated with a multivariate rank test. Among different versions of multivariate median and HL-type estimates of location, the vector of coordinatewise median and the vector of coordinatewise HL estimates are perhaps the simplest ones. In one sample multivariate location problems, the tests that are naturally associated with them are the coordinatewise sign tests and the coordinatewise signed rank tests respectively. Both the tests have been studied extensively by

Bickel (1965), Puri and Sen (1971) etc. Spatial median is another popular generalization of univariate median to multidimension, which is considered in earlier chapters and the test that is naturally associated with spatial median is the angle test. The angle test, based on the direction vectors $\boldsymbol{U}(\boldsymbol{X}_i) = \|\boldsymbol{X}_i\|^{-1}\boldsymbol{X}_i$ $(1 \leq i \leq n, \boldsymbol{X}_i \neq 0)$, has been considered by Brown(1983, 1988), Möttönen and Oja (1995) etc..

One serious drawback of coordinatewise sign as well as signed rank test is that neither of them is invariant under arbitrary affine transformation of the data. In addition to being an undesirable geometric feature, this lack of invariance is known to have some negative impact on the statistical performance of these tests especially when the real-valued components of the multivariate data are substantially correlated. This issue was first raised by Bickel (1964,1965), and subsequently investigated by Brown and Hettmansperger (1987,1989). Spatial median is known to have rather impressive and somewhat counter-intuitive efficiency properties for multidimensional data generated from a spherically symmetric probability distribution, and this has been discussed in detail in Chaudhuri (1992a) [see also Brown (1983)]. However, the performance of angle test tends to be quite poor compared to other affine invariant procedures when there is significant deviation from spherical symmetry caused by the presence of correlation among observed variables (e.g. when the underlying distribution is elliptically symmetric). While the coordinatewise sign and signed rank tests can be used for data consisting of variables measured in different scales, it is not possible to use angle test on such data due to its lack of invariance under coordinatewise scale transformations.

In Section 3.2, we propose the transformation technique for constructing affine invariant sign and rank tests in one sample problems. In the same section, we will see that our tests have a very encouraging common feature – they inherit good efficiency properties of coordinatewise sign and signed rank tests in spherically symmetric multivariate normal models and extend them to more general elliptically symmetric situations. In Section 3.3, we will discuss the two sample tests and their asymptotic properties. Statistical performance of the proposed tests under various elliptically symmetric distributions is studied with the help of simulation in Section 3.4. In the same Section, we will indicate how one can estimate the P-value when our proposed tests are applied to real data sets by simulating the permutation distributions of the proposed test statistics under the null hypothesis.

3.2 One Sample Location Problem

Suppose that we have n data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$, and assume that $n > d+1$. Let us begin by introducing some notation. As in the preceding Chapter define

$$S_n = \{\alpha | \alpha \subseteq \{1, 2, \ldots, n\} \text{ and } \#\{i : i \in \alpha\} = d+1\},$$

which is the collection of all subsets of size $d+1$ of $\{1, 2, \ldots, n\}$. For a fixed subset $\alpha = \{i_0, i_1, \ldots, i_d\}$, consider the points $\boldsymbol{X}_{i_0}, \boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_d}$, which will form a 'data driven coordinate system' as described before, and the $d \times d$ matrix $\mathbf{X}(\alpha)$ containing the columns $\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{i_d} - \boldsymbol{X}_{i_0}$ can be taken as the transformation matrix for transforming the data points $\boldsymbol{X}_j$ such that $1 \leq j \leq n$, $j \notin \alpha$ in order to express them in terms of the new coordinate system as $\boldsymbol{Y}_j^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_j$. If the observations $\boldsymbol{X}_i$'s are generated from a common distribution which is absolutely continuous w.r.t the Lebesgue measure on $\mathbb{R}^d$, the transformation matrix $\mathbf{X}(\alpha)$ must be an invertible matrix with probability one. We have already observed that the transformed observations $\mathbf{Y}_j^{(\alpha)}$'s with $1 \leq j \leq n$ and $i \notin \alpha$ form a maximal invariant with respect to the group of nonsingular linear transformations on $\mathbb{R}^d$ [see Chapter 2].

3.2.1 Affine Invariant Sign and Signed Rank Tests

One can now define the following test statistics

$$T_n^{(\alpha)} = \sum_{i \notin \alpha} Sign(\boldsymbol{Y}_i^{(\alpha)}) \tag{3.1}$$

and

$$R_n^{(\alpha)} = \frac{1}{n} \sum_{i \notin \alpha} \sum_{\substack{j \notin \alpha \\ j \neq i}} Sign(\boldsymbol{Y}_i^{(\alpha)} + \boldsymbol{Y}_j^{(\alpha)}), \tag{3.2}$$

where $Sign(\boldsymbol{X}_i) = (Sign(X_{i1}), \ldots, Sign(X_{id}))^T$ for $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$. In view of the construction, it is clear that $T_n^{(\alpha)}$ and $R_n^{(\alpha)}$ are affine invariant. Note that they are nothing but coordinatewise sign test statistic and coordinatewise Wilcoxon's signed rank statistic respectively based on the transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s. A question that naturally arises at this stage is how to choose the 'data driven coordinate system' or equivalently the data based transformation matrix $\mathbf{X}(\alpha)$. An answer to this question is provided in the following discussion.

Let us consider the elliptically symmetric density $h(\boldsymbol{x}) = \{\det(\Sigma)\}^{-1/2} f(\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x})$, where Σ is a $d \times d$ positive definite matrix, and $f(\boldsymbol{x}^T \boldsymbol{x})$ is a continuous spherically symmetric density around the origin in $\mathbb{R}^d$. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are assumed to be i.i.d observations

with common elliptically symmetric density $h(\boldsymbol{x}-\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the location of elliptic symmetry for the data. Suppose that we have two competing hypothesis $H_0 : \boldsymbol{\theta} = 0$ and $H_A : \boldsymbol{\theta} \neq 0$. We now state a result that summarizes the main features of the affine invariant sign test statistic $T_n^{(\alpha)}$.

Theorem 3.2.1 *Under the null hypothesis $H_0 : \boldsymbol{\theta} = 0$, the conditional distribution of $n^{-1/2}T_n^{(\alpha)}$ given the $\boldsymbol{X}_i$'s with $i \in \alpha$ does not depend on f, and it converges to d-variate normal with zero mean and a variance covariance matrix $\Psi_1\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}$ as $n \longrightarrow \infty$, where Ψ_1 depends only on $\Sigma^{-1/2}\mathbf{X}(\alpha)$. When $\log f$ is twice differentiable almost everywhere (w.r.t. Lebesgue measure) on $\mathbb{R}^d$ and satisfies the Cramér type regularity conditions, the alternatives $H_A^{(n)} : \boldsymbol{\theta} = n^{-1/2}\boldsymbol{\delta}$ such that $\boldsymbol{\delta} \in \mathbb{R}^d$ and $\boldsymbol{\delta} \neq \mathbf{0}$ will form a contiguous sequence, and the conditional limiting distribution of $n^{-1/2}T_n^{(\alpha)}$ under that sequence of alternatives is normal with the same variance covariance matrix Ψ_1 and a mean vector $\Lambda_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$ that depends on f, $\Sigma^{-1/2}\boldsymbol{\delta}$ and $\Sigma^{-1/2}\mathbf{X}(\alpha)$. Also, the limiting conditional power of the test under such a sequence of contiguous alternatives increases monotonically with the noncentrality parameter*

$$\Delta_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$$

$$= [\Lambda_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\alpha)\}]^T[\Psi_1\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1}\Lambda_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma\mathbf{X}(\alpha)\}. \qquad (3.3)$$

Here the noncentrality parameter Δ_1 is such that for any f, $\boldsymbol{\delta}$, Σ and any invertible matrix $\mathbf{A}$, we have $\Delta_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{A}\} = \beta(f)\boldsymbol{\delta}^T\Sigma^{-1}\boldsymbol{\delta}$, where β is a scalar depending only on f whenever $\mathbf{A}^T\Sigma^{-1}\mathbf{A}$ is a diagonal matrix. Further, for any invertible matrix $\mathbf{B}$ we will have

$$\inf_{\delta:\delta^T\Sigma^{-1}\delta=c} \Delta_1\{f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{B}\} \leq c\beta(f).$$

Proof: First note that in view of the affine invariance of the test statistic $T_n^{(\alpha)}$, it is enough to prove the Theorem only for $\Sigma = \mathbf{I}_d$. Now, since given $\mathbf{X}(\alpha)$, the transformed observations $\boldsymbol{Y}_i^{(\alpha)}$'s are conditionally independent, and they are identically distributed with the elliptically symmetric density $|\det\{\mathbf{X}(\alpha)\}|f[\boldsymbol{y}^T\{\mathbf{X}(\alpha)\}^T\mathbf{X}(\alpha)\boldsymbol{y}]$, the conditional distribution of $T_n^{(\alpha)}$ does not depend on f and depends only on $\mathbf{X}(\alpha)$. This actually follows from the introductory discussion in Bickel (1965). The (i,j)-th element of the variance covariance matrix $\Psi_1\{\mathbf{X}(\alpha)\}$ is $r_i r_j(2/\pi)\sin^{-1}\rho_{ij}$ where $\{\mathbf{X}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$, $\mathbf{R}(\alpha) = diag(r_1, \ldots, r_d)$ and each row of the matrix $\mathbf{J}(\alpha)$ is of unit length. ρ_{ij} is the inner product between the i-th and j-th row of $\mathbf{J}(\alpha)$.

When $\log f$ is twice differentiable almost everywhere in $\mathbb{R}^d$ and satisfies Cramér type regularity conditions, by Lemma 3.1 of Bickel (1965), the conditional limiting distribution of $n^{-1/2}T_n^{(\alpha)}$ given $\mathbf{X}(\alpha)$ under the sequence of contiguous alternatives $H_A^{(n)}$ is normal with

mean equal to $\{\beta(f)\}^{1/2}\{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{\delta} = \Lambda_1\{f, \boldsymbol{\delta}, \mathbf{X}(\alpha)\}$ and Ψ_1 as the variance covariance matrix. Here $\beta(f) = \{2f_1(0)\}^2$ is a scalar multiple that depends only on the univariate marginal f_1 of the spherically symmetric density f. This immediately implies that the conditional limiting distribution of $n^{-1}\{T_n^{(\alpha)}\}^T[\Psi_1\{\mathbf{X}(\alpha)\}]^{-1}T_n^{(\alpha)}$ under $H_A^{(n)}$ is noncentral χ^2 with d d.f., and the noncentrality parameter is

$$\beta(f)\boldsymbol{\delta}^T[\{\mathbf{J}(\alpha)\}^{-1}\mathbf{D}(\alpha)[\{\mathbf{J}(\alpha)\}^T]^{-1}]^{-1}\boldsymbol{\delta}, \tag{3.4}$$

where the (i,j)-th element of the matrix $\mathbf{D}(\alpha)$ is $(2/\pi)\sin^{-1}\rho_{ij}$. Consequently the limiting conditional power of the test under the sequence of contiguous alternatives will be a monotonically increasing function of this noncentrality parameter. Note now that when the rows of $\mathbf{J}(\alpha)$ are orthogonal we have $\{\mathbf{J}(\alpha)\}^{-1}\mathbf{D}(\alpha)[\{\mathbf{J}(\alpha)\}^T]^{-1} = \mathbf{I}_d$. Finally the minimax ordering of the noncentrality parameter Δ_1 stated in the Theorem follows from Theorem 2.2.3 of Chapter 2 which implies that the largest eigen value of $\{\mathbf{J}(\alpha)\}^{-1}\mathbf{D}(\alpha)[\{\mathbf{J}(\alpha)\}^T]^{-1}$ is larger than or equal to 1. □

The main implication of the above Theorem is that whatever may f be, it is possible to simulate the conditional finite sample null distribution of $T_n^{(\alpha)}$ after obtaining an appropriate estimate of Σ in small sample situations, and one can use normal approximation when the sample size is large. It is interesting to note that in order to maximize the minimum power of the test one needs to choose $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix (especially for alternatives close to the null). We now make the following observation, from the proof of the above Theorem.

Observation 2.1 : The Pitman efficiency of the test based on $T_n^{(\alpha)}$ with the above choice of $\mathbf{X}(\alpha)$ would be close to that of coordinatewise sign test when the coordinate variables are uncorrelated (i.e. Σ is diagonal). Further, it will outperform coordinatewise sign test in elliptically symmetric models in a minimax sense i.e. if the parameter $\boldsymbol{\theta}$ is non-zero and oriented in the worst possible direction giving rise to the minimum power of the tests, our invariant test will have more power than the noninvariant sign test. Also, this version of affine invariant sign test will be more (or less) efficient than the usual Hotelling's T^2-test if the tail of the density f is 'heavy' (or 'light').

Next we state a result on the behavior of affine invariant signed rank test statistic $R_n^{(\alpha)}$ defined in (3.2).

Theorem 3.2.2 *As $n \longrightarrow \infty$, under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$, the conditional distribution of $n^{-1/2}R_n^{(\alpha)}$ given the $\boldsymbol{X}_i$'s with $i \in \alpha$ converges to a d-variate normal distribution with zero mean and a variance covariance matrix $\Psi_2\{f, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$ that depends on f and $\Sigma^{-1/2}\mathbf{X}(\alpha)$. When $\log f$ is twice differentiable almost everywhere (w.r.t. Lebesgue*

measure) on $\mathbb{R}^d$ *and satisfies Cramér type regularity conditions, the conditional limiting distribution of* $n^{-1/2}R_n^{(\alpha)}$ *under the sequence of contiguous alternatives* $H_A^{(n)} : \theta = n^{-1/2}\delta$ *such that* $\delta \in \mathbb{R}^d$ *and* $\delta \neq 0$ *will be normal with the same variance covariance matrix* Ψ_2 *and a mean vector* $\Lambda_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$ *that depends on* f, $\Sigma^{-1/2}\delta$ *and* $\Sigma^{-1/2}\mathbf{X}(\alpha)$. *Also, the limiting conditional power of the test under such a sequence of alternatives depends monotonically on the noncentrality parameter*

$$\Delta_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$$

$$= [\Lambda_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}]^T[\Psi_2\{f, \Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1}\Lambda_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}. \quad (3.5)$$

Here the noncentrality parameter Δ_2 *is such that for any* f, δ, Σ *and any invertible matrix* $\mathbf{A}$, *we have* $\Delta_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{A}\} = \gamma(f)\delta^T\Sigma^{-1}\delta$, *where* γ *is a scalar depending only on* f, *whenever* $\mathbf{A}^T\Sigma^{-1}\mathbf{A}$ *is a diagonal matrix. Further, normality of the density* f *implies that for any invertible matrix* $\mathbf{B}$, *we have*

$$\inf_{\delta:\delta^T\Sigma^{-1}\delta=c} \Delta_2\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{B}\} \leq c\gamma(f).$$

Proof: Here also we can assume that $\Sigma = \mathbf{I}_d$ in view of the affine invariance of the test statistic $R_n^{(\alpha)}$. The conditional limiting distribution of $n^{-1/2}R_n^{(\alpha)}$ under the null hypothesis and also under the sequence of contiguous alternatives follow from Bickel (1965). Let F_i be the marginal distribution of the i-th coordinate of the transformed observations $\mathbf{Y}_j^{(\alpha)}$'s. Then the (i,j)-th element of the variance covariance matrix Ψ_2 is given by the covariance between $F_i(Y_{ki})$ and $F_j(Y_{kj})$, where Y_{ki} is the i-th coordinate of $\mathbf{Y}_k^{(\alpha)}$. And the noncentrality parameter of the limiting distribution under alternative $H_A^{(n)}$ is given by $\delta^T[\mathcal{V}\{f, \mathbf{I}_d, \mathbf{X}(\alpha)\}]^{-1}\delta$, where $\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\}$ is as defined in (2.2). Consequently, the minimax ordering of the noncentrality parameter under the assumption of normality of the density f follows from arguments similar to those used in the proof of Theorem 3.2.1 and Theorem 2.2.3 of Chapter 2. □

From Theorem 3.2.2, it is clear that when the underlying distribution is normal, the optimal choice for $\mathbf{X}(\alpha)$ is such that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ is as close as possible to a diagonal matrix in order to maximize the minimum power of the test. We now observe the following.

Observation 2.2 : The asymptotic Pitman efficiency of the test based on $R_n^{(\alpha)}$ can be made close to that of coordinatewise Wilcoxon's signed rank test when Σ is diagonal by selecting $\mathbf{X}(\alpha)$ as described above, and it will outperform the coordinatewise signed rank test in a minimax sense when Σ is not a diagonal matrix and data follows normal distribution.

Whether the underlying distribution is normal or not, it is easy to simulate the null permutation distribution of $R_n^{(\alpha)}$ and based on that one can carry out the permutation test. Furthermore, one can use normal approximation to the null distribution of $R_n^{(\alpha)}$ for large values of n. Later, we will compare the finite sample performance of our affine invariant test statistics $T_n^{(\alpha)}$ and $R_n^{(\alpha)}$ with some other standard one sample tests for multivariate location.

3.2.2 An Affine Invariant Multivariate Angle Test

As in the preceding Subsection, let us assume that the X_i's are i.i.d observations generated from the elliptically symmetric density $h(x-\theta)$ on $\mathbb{R}^d$. Suppose that we have two competing hypotheses $H_0 : \theta = 0$ and $H_A : \theta \neq 0$ concerning the center of elliptic symmetry of the distribution. Consider once again the transformed observations $Y_j^{(\alpha)}$ for $j \notin \alpha$ and $1 \leq j \leq n$, and define the test statistic $\mathbf{U}_n^{(\alpha)} = \sum_{j \notin \alpha} \|Y_j^{(\alpha)}\|^{-1} Y_j^{(\alpha)}$. We now state a Theorem that summarizes the main features of this test statistic.

Theorem 3.2.3 *Under the null hypothesis $H_0 : \theta = 0$, the conditional distribution of $\mathbf{U}_n^{(\alpha)}$ given the X_i's with $i \in \alpha$ does not depend on f, and it depends on Σ through $\Sigma^{-1/2}\mathbf{X}(\alpha)$. Further, in large samples, the conditional null distribution of $n^{-1/2}\mathbf{U}_n^{(\alpha)}$ is approximately normal with zero mean and a variance covariance matrix $\Psi\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}$ that depends on $\Sigma^{-1/2}\mathbf{X}(\alpha)$. When $\log f$ is twice differentiable almost everywhere (w.r.t. Lebesgue measure) on $\mathbb{R}^d$ and satisfies the Cramer type regularity conditions, the alternatives $H_A^{(n)} : \theta = n^{-1/2}\delta$ such that $\delta \in \mathbb{R}^d$ and $\delta \neq 0$ will form a contiguous sequence, and the conditional limiting distribution of $n^{-1/2}\mathbf{U}_n^{(\alpha)}$ under that sequence of alternatives is normal with the same variance covariance matrix Ψ and a mean vector $\Lambda\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$ that depends on f, $\Sigma^{-1/2}\delta$ and $\Sigma^{-1/2}\mathbf{X}(\alpha)$. Also, the limiting conditional power of the test under such a sequence of contiguous alternatives depends monotonically on the noncentrality parameter*

$$\delta\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}$$

$$= [\Lambda\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}]^T [\Psi\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1} \Lambda\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{X}(\alpha)\}\ ,$$

where δ is such that for any f, δ, Σ and any two invertible matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\delta\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{B}\} \geq \delta\{f, \Sigma^{-1/2}\delta, \Sigma^{-1/2}\mathbf{A}\}$ whenever $\mathbf{B}^T\Sigma^{-1}\mathbf{B} = \lambda \mathbf{I}_d$ for some $\lambda > 0$.

Proof: First note that in view of the affine invariance of the test statistic $\mathbf{U}_n^{(\alpha)}$, it is enough to prove the entire Theorem only for $\Sigma = \mathbf{I}_d$. Now, since given $\mathbf{X}(\alpha)$ the

transformed observations $\boldsymbol{Y}_j^{(\alpha)}$'s are conditionally independent and they are identically distributed with the elliptically symmetric density $|\det\{\mathbf{X}(\alpha)\}| f[y^T\{\mathbf{X}(\alpha)\}^T\mathbf{X}(\alpha)y]$, the conditional distribution of $\mathbf{U}_n^{(\alpha)} = \sum_{j\notin\alpha} \|\boldsymbol{Y}_j^{(\alpha)}\|^{-1}\boldsymbol{Y}_j^{(\alpha)}$ does not depend on f and depends only on $\mathbf{X}(\alpha)$. This actually follows from what we have already seen in the proof of Theorem 2.3.1. Next, the asymptotic normality of the conditional null distribution of $n^{-1/2}\mathbf{U}_n^{(\alpha)}$ follows by a straight forward application of the central limit theorem, and the variance covariance matrix Ψ is equal to the matrix $\mathbf{C}$ defined in the proof of Theorem 2.3.1.

Some standard analysis using Le Cam's third lemma [see Hajek and Sidak (1967)] and the spherical symmetry of the density f imply that under the sequence $H_A^{(n)}$, the conditional limiting distribution of $n^{-1/2}\mathbf{U}_n^{(\alpha)}$ given $\mathbf{X}(\alpha)$ is normal with mean equal to $\beta(d,f)\mathbf{H}\{\mathbf{X}(\alpha)\}\delta = \Lambda\{f,\delta,\mathbf{X}(\alpha)\}$ and $\Psi\{\mathbf{X}(\alpha)\}$ as the variance covariance matrix. Here β is a scalar multiple that depends only on the dimension d and the density f, and the $d\times d$ matrix $\mathbf{H}\{\mathbf{X}(\alpha)\}$ is equal to $\boldsymbol{COV}[\|\{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{Y}\|^{-1}\{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{Y}\ ,\ \|\boldsymbol{Y}\|^{-1}\boldsymbol{Y}]$, where $\boldsymbol{Y}$ is a d-dimensional random vector with density $f(y^Ty)$. This immediately implies that the conditional limiting distribution of $n^{-1}\{\mathbf{U}_n^{(\alpha)}\}^T[\Psi\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1}\mathbf{U}_n^{(\alpha)}$ under $H_A^{(n)}$ is noncentral χ^2 with d d.f. and noncentrality parameter $\delta\{f,\delta,\mathbf{X}(\alpha)\}$, which is defined in the statement of the Theorem. Consequently the limiting conditional power of the test under the sequence of contiguous alternatives will be a monotonically increasing function of this δ. Finally the ordering of δ stated in the Theorem will follow if we can show that

$$\{\mathbf{H}(\mathbf{A})\}^T\{\Psi(\mathbf{A})\}^{-1}\mathbf{H}(\mathbf{A}) \leq_{n.n.d} \{\mathbf{H}(\mathbf{B})\}^T\{\Psi(\mathbf{B})\}^{-1}\mathbf{H}(\mathbf{B})\ ,$$

for any two nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{B}^T\mathbf{B} = \lambda\mathbf{I}_d$. The proof of this nonnegative definite ordering of matrices follows from straightforward arguments that are very similar to those used in the second half of the proof of Theorem 2.3.1. □

Note that one of the main implications of this Theorem is that whatever may f be, it is possible to simulate the conditional finite sample null distribution of $\mathbf{U}_n^{(\alpha)}$ after obtaining an appropriate estimate of Σ in small sample situations, and one can use the normal approximation when the sample size is large. It is also noteworthy that in order to maximize the power of the test that rejects H_0 for large values of $\{\mathbf{U}_n^{(\alpha)}\}^T[\Psi\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1}\mathbf{U}_n^{(\alpha)}$, one needs to choose $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix with all diagonal entries equal (especially for alternatives close to the null). It is transparent from the proof of Theorem 3.2.3 that by choosing $[\mathbf{X}(\alpha)]^T\Sigma^{-1}\mathbf{X}(\alpha)$ very close to a matrix of the form $\lambda\mathbf{I}_d$, asymptotic Pitman efficiency of the test can be made close to that of the angle test in spherically symmetric model (i.e. when $\Sigma = \sigma^2\mathbf{I}_d$), and it will be more efficient than angle test in elliptically symmetric models. Also, the test will be more (or less) efficient than the standard test of location based on Hotelling's

T^2 statistic (which is a test based on the sample mean vector) if the tail of the density f is 'heavy' (or 'light'). In the Section 3.4, we will demonstrate how the simulated conditional null distribution of $\mathbf{U}_n^{(\alpha)}$ can be used to determine the critical region of the test for a pre-specified level of significance and also to estimate the P-value corresponding to a given value of $\{\mathbf{U}_n^{(\alpha)}\}^T[\Psi\{\Sigma^{-1/2}\mathbf{X}(\alpha)\}]^{-1}\mathbf{U}_n^{(\alpha)}$ computed from the observed data. Also, there we will compare the finite sample performance of our test with some other standard one sample tests for multivariate location.

3.3 Two Sample Location Problem

Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m$ and $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ be two independent samples from two d-dimensional distributions For the purpose of comparing these two samples, we will develop an affine invariant multivariate analog of Wilcoxon's two-sample rank-sum test based on transformation and retransformation approach. Let us define $\boldsymbol{X}_{m+1}=\boldsymbol{Y}_1,\ldots,\boldsymbol{X}_{m+n}=\boldsymbol{Y}_n$ and the transformation matrix $\mathbf{X}(\alpha)$ whose columns are $\boldsymbol{X}_{i_1}-\boldsymbol{X}_{i_1^*},\ldots,\boldsymbol{X}_{i_d}-\boldsymbol{X}_{i_d^*}$, where $\alpha=\{j_1,j_2,i_1,\ldots,i_d\}\in S_{m+n}$ with $1\le j_1\le m$, $m+1\le j_2\le m+n$ and $i_k^*=j_1$ or j_2 depending on whether $i_k\in\{1,\ldots,m\}$ or $i_k\in\{m+1,\ldots,m+n\}$ respectively for all $1\le k\le d$. As before, define transformed observations $\boldsymbol{Z}_j^{(\alpha)}=\{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_j$, $1\le j\le m+n$, $j\notin\alpha$. One can now compute coordinatewise Wilcoxon's rank-sum statistic based on $\boldsymbol{Z}_j^{(\alpha)}$'s, which is equivalent to the statistic

$$W_{m,n}^{(\alpha)}=\frac{1}{m+n}\sum_{\substack{i=1\\ i\notin\alpha}}^{m}\sum_{\substack{j=1\\ m+j\notin\alpha}}^{n}Sign(\boldsymbol{Z}_i^{(\alpha)}-\boldsymbol{Z}_{m+j}^{(\alpha)}). \tag{3.6}$$

The following result describes the main features of the test statistic $W_{m,n}^{(\alpha)}$. Let $\boldsymbol{X}_i$'s be i.i.d. with density $h(\boldsymbol{x}-\boldsymbol{\theta})$, $\boldsymbol{Y}_i$'s be i.i.d. with density $h(\boldsymbol{x}-\boldsymbol{\theta}-\boldsymbol{\delta})$, where h is as before, and we want to test $H_0:\boldsymbol{\delta}=0$ against $H_A:\boldsymbol{\delta}\neq 0$.

Theorem 3.3.1 *Under the null hypothesis $H_0:\boldsymbol{\delta}=0$, the conditional distribution of $(m+n)^{-1/2}W_{m,n}^{(\alpha)}$ given the $\boldsymbol{X}_i$'s with $i\in\alpha$ converges to a normal distribution with zero mean and variance covariance matrix $\Psi_3\{f,\Sigma^{-1/2}\mathbf{X}(\alpha),\lambda\}$ as $m,n\longrightarrow\infty$ such that $m/(m+n)\to\lambda>0$. When $\log f$ is twice differentiable almost everywhere (w.r.t. Lebesgue measure on $\mathbb{R}^d$) and satisfies Cramér type regularity conditions, the conditional limiting distribution of $(m+n)^{-1/2}W_{m,n}^{(\alpha)}$ under the sequence of contiguous alternatives $H_A^{(m,n)}:\boldsymbol{\delta}=(m+n)^{-1/2}\boldsymbol{\mu}$ such that $\boldsymbol{\mu}\in\mathbb{R}^d$ and $\boldsymbol{\mu}\neq 0$ is normal with the same variance covariance matrix Ψ_3 and a mean vector $\Lambda_3\{f,\Sigma^{-1/2}\boldsymbol{\mu},\Sigma^{-1/2}\mathbf{X}(\alpha),\lambda\}$ that depends on f, $\Sigma^{-1/2}\boldsymbol{\mu}$ and $\Sigma^{-1/2}\mathbf{X}(\alpha)$. Also, the limiting conditional power of the test under such a sequence of*

alternatives increases monotonically with the noncentrality parameter

$$\Delta_3\{f, \Sigma^{-1/2}\mu, \Sigma^{-1/2}\mathbf{X}(\alpha), \lambda\}$$

$$= [\Lambda_3\{f, \Sigma^{-1/2}\mu, \Sigma^{-1/2}\mathbf{X}(\alpha), \lambda\}]^T[\Psi_3\{f, \Sigma^{-1/2}\mathbf{X}(\alpha), \lambda\}]^{-1}\Lambda_3\{f, \Sigma^{-1/2}\mu, \Sigma^{-1/2}\mathbf{X}(\alpha), \lambda\}. \tag{3.7}$$

Further, for any μ, Σ and any invertible matrix $\mathbf{A}$ such that $\mathbf{A}^T\Sigma^{-1}\mathbf{A}$ is a diagonal matrix, we have $\Delta_3\{f, \Sigma^{-1/2}\mu, \Sigma^{-1/2}\mathbf{B}, \lambda\} = \eta(f, \lambda)\mu^T\Sigma^{-1}\mu$, where the scalar η depends only on f and λ, and for normal f and any invertible matrix $\mathbf{B}$, we have

$$\inf_{\mu:\mu^T\Sigma^{-1}\mu=c} \Delta_3\{f, \Sigma^{-1/2}\mu, \Sigma^{-1/2}\mathbf{B}, \lambda\} \leq c\eta(f, \lambda).$$

Proof: In view of the affine invariance of the test statistic $W_{m,n}^{(\alpha)}$, we may assume that $\Sigma = \mathbf{I}_d$. Under H_0, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are assumed to have the same distribution with density function $h(\boldsymbol{x} - \boldsymbol{\theta})$. We can generalize the result of asymptotic normality of the univariate two sample rank-sum test statistic to the d-dimensional rank statistic $W_{m,n}^{(\alpha)}$ through U-statistics type representations [for a detailed proof, see Hodges and Lehmann (1963), Puri and Sen (1971)]. Let F_i be the marginal distribution of the i-th coordinate of the transformed observations $\boldsymbol{Z}_j^{(\alpha)}$'s and $m/(m+n) \longrightarrow \lambda$ as $m, n \longrightarrow \infty$. Then the (i, j)-th element of the asymptotic variance covariance matrix of $(m+n)^{-1/2}W_{m,n}^{(\alpha)}$ is given by $\lambda(1-\lambda)Cov\{F_i(Z_{ki}), F_j(Z_{kj})\}$, where Z_{ki} is the i-th coordinate of $\boldsymbol{Z}_k^{(\alpha)}$.

When $\log f$ is twice differentiable almost everywhere in $\mathbb{R}^d$ and satisfies Cramér type regularity conditions, the sequence of alternatives $H_A^{(m,n)} : \delta = (m+n)^{-1/2}\mu$ such that $\mu \in \mathbb{R}^d$ and $\mu \neq 0$ will form a contiguous sequence. Then we can apply results of Hajek and Sidak (1967) for univariate two sample linear rank statistic to have the asymptotic normality of $(m+n)^{-1/2}W_{m,n}^{(\alpha)}$ given $\mathbf{X}(\alpha)$ under the sequence of contiguous alternatives $H_A^{(m,n)}$ as $m, n \longrightarrow \infty$. The asymptotic mean of the non-null distribution is given by $\lambda(1-\lambda)\{\int f_1^2(x)\,dx\}\mathbf{J}(\alpha)\mu$ where the matrix $\mathbf{J}(\alpha)$ and the function f_1 are as defined earlier. Consequently, the noncentrality parameter of the limiting distribution under the alternative $H_A^{(m,n)}$ is given by $\lambda(1-\lambda)\mu^T[\mathcal{V}\{f, \mathbf{I}_d, \mathbf{X}(\alpha)\}]^{-1}\mu$ where $\mathcal{V}\{f, \Sigma, \mathbf{X}(\alpha)\}$ is as defined in (2.8). Again as before the minimax ordering of the noncentrality parameter under the assumption of normality of the density f follows from the arguments used in the proofs of Theorems 3.2.1 and 3.2.2 and Theorem 2.2.3 of Chapter 2. □

Once again for normal f, if one chooses the transformation matrix $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ is as close as possible to a diagonal matrix, it maximizes the minimum power of the test. We now note the following.

Observation 2.3 : With proper choice of $\mathbf{X}(\alpha)$, the Pitman efficiency of the test based on $W_{m,n}^{(\alpha)}$ can be made close to that of coordinatewise two sample rank-sum test when Σ is diagonal. Also, if the data follows normal distribution, it will outperform the coordinatewise two sample rank-sum test in a minimax sense when Σ is not diagonal.

We close this section by pointing out that using a similar approach one may define an affine invariant version of multivariate median test in the two sample problem. Define

$$V_{m,n}^{(\alpha)} = \sum_{\substack{i=1 \\ i \notin \alpha}}^{m} Sign(Z_i^{(\alpha)} - \hat{\phi}), \tag{3.8}$$

where $\hat{\phi}$ is the vector of coordinatewise medians based on the transformed observations of the combined sample $\boldsymbol{Z}_i^{(\alpha)}$, $i = 1, \ldots, m+n$, $i \notin \alpha$. It is possible to establish the asymptotic normality of $(m+n)^{-1/2} V_{m,n}^{(\alpha)}$ following the results in Puri and Sen (1971). Here also we may adopt the strategy for selecting the transformation matrix $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T \Sigma^{-1} \mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix. In the next section, we will explain in detail how to obtain P-values for the test and report simulation results comparing our test with other standard multivariate two sample tests for different multivariate distributions.

3.4 Simulation Results and Data Analysis

From the results and discussions in the previous section, it is clear that we need to choose the transformation matrix $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T \Sigma^{-1} \mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix. Since Σ will be unknown in practice, we have to estimate that from the data, and we will need a consistent estimate that will be invariant under location shift and equivariant under linear transformations of the data (say $\hat{\Sigma}$). If we can assume that the underlying distribution has finite second order moments, we can use the usual variance covariance matrix for this purpose. To be in conformity with the selection procedure described in Chapter 2, we will try to select $\mathbf{X}(\alpha)$ in such a way that the eigenvalues of the positive definite matrix $\{\mathbf{X}(\alpha)\}^T \hat{\Sigma}^{-1} \mathbf{X}(\alpha)$ become as equal as possible. Rather than computing the eigenvalues of the matrix explicitly, we will minimize the ratio between the arithmetic mean and the geometric mean of the eigenvalues, which are given by the trace and the determinant of the matrix respectively. We have observed that instead of minimizing the ratio over all possible subsets α with size $d+1$ of $\{1, \ldots, n\}$, one can substantially reduce the amount of computation by stopping the search for optimal subset α as soon as the ratio becomes sufficiently close to one. Of course there are other different ways to achieve this goal of making $\{\mathbf{X}(\alpha)\}^T \hat{\Sigma}^{-1} \mathbf{X}(\alpha)$ as close as possible to a diagonal

matrix. We have adopted a technique that is computationally convenient and has been observed to work fairly well in our numerical investigations.

3.4.1 Simulated Powers of Different One Sample Tests

We will present the results of a simulation study that was carried out to compare the finite sample powers of our affine invariant rank test, sign test and angle test, which are based on the statistics $R_n^{(\alpha)}$, $T_n^{(\alpha)}$ and $\mathbf{U}_n^{(\alpha)}$ respectively in the one sample problem

Table 3.1: *Finite sample power of affine invariant rank test and its competitors in the one sample problem for $n = 30$, $d = 2$ and level of significance = 5%.*

Distributions	Test	ρ	$(\theta^T\Sigma^{-1}\theta)^{1/2}$					
			0.0	0.3	0.6	0.9	1.2	1.5
Laplace	$R_n^{(\alpha)}$	-	0.049	0.134	0.405	0.729	0.918	0.982
	$T_n^{(\alpha)}$	-	0.051	0.134	0.384	0.677	0.891	0.970
	$\mathbf{U}_n^{(\alpha)}$	-	0.050	0.149	0.421	0.740	0.923	0.982
	T^2	-	0.049	0.121	0.367	0.692	0.902	0.982
	S_n	0.00	0.048	0.133	0.383	0.677	0.889	0.972
		0.75	0.047	0.134	0.413	0.708	0.905	0.971
		0.85	0.046	0.133	0.419	0.712	0.892	0.946
		0.95	0.035	0.123	0.392	0.647	0.796	0.818
Normal	$R_n^{(\alpha)}$	-	0.048	0.266	0.796	0.986	0.990	0.998
	$T_n^{(\alpha)}$	-	0.053	0.193	0.615	0.923	0.994	0.999
	$\mathbf{U}_n^{(\alpha)}$	-	0.049	0.207	0.679	0.954	0.998	1.000
	T^2	-	0.049	0.267	0.805	0.989	1.000	1.000
	S_n	0.00	0.049	0.184	0.609	0.925	0.994	1.000
		0.75	0.046	0.196	0.645	0.938	0.984	0.969
		0.85	0.049	0.194	0.648	0.927	0.947	0.899
		0.95	0.033	0.175	0.597	0.831	0.782	0.667
t with 3 d.f.	$R_n^{(\alpha)}$	-	0.051	0.191	0.580	0.876	0.977	0.983
	$T_n^{(\alpha)}$	-	0.050	0.172	0.523	0.848	0.969	0.995
	$\mathbf{U}_n^{(\alpha)}$	-	0.051	0.178	0.553	0.883	0.988	0.999
	T^2	-	0.043	0.157	0.489	0.799	0.938	0.981
	S_n	0.00	0.052	0.174	0.528	0.844	0.970	0.996
		0.75	0.048	0.184	0.559	0.868	0.973	0.980
		0.85	0.045	0.184	0.564	0.865	0.949	0.931
		0.95	0.035	0.169	0.524	0.778	0.819	0.749

with the powers of the well-known Hotelling's T^2 test and the noninvariant sign test,

which is based on the coordinatewise sign test statistic S_n. We have used sample size $n = 30$, and for level of significance 5%, we estimated the powers in each case from 5,000 Monte Carlo replications for $d = 2$ and from 3,000 Monte Carlo replications for $d = 3$. We generated data from three different distributions, namely multivariate normal, multivariate Laplace (i.e. when $f(x^T x) = k \exp\{-(x^T x)^{1/2}\})$ and multivariate t with 3 degrees of freedom and for Σ we have used the matrix with each diagonal entry equal to 1 and each off diagonal entry equal to ρ such that $\rho \in [0, 1)$. The results are presented in Tables 3.1 and 3.2. For Hotelling's T^2 test the critical value at 5% level of significance

Table 3.2: Finite sample power of affine invariant rank test and its competitors in the one sample problem for $n = 30$, $d = 3$ and level of significance = 5%.

Distributions	Test	ρ	$(\theta^T \Sigma^{-1} \theta)^{1/2}$					
			0.0	*0.3*	*0.6*	*0.9*	*1.2*	*1.5*
Laplace	$R_n^{(a)}$	-	0.052	0.103	0.272	0.530	0.775	0.922
	$T_n^{(a)}$	-	0.051	0.103	0.247	0.471	0.694	0.857
	$\mathbf{U}_n^{(a)}$	-	0.051	0.104	0.258	0.497	0.757	0.916
	T^2	-	0.042	0.083	0.237	0.493	0.740	0.911
	S_n	0.00	0.046	0.080	0.204	0.415	0.662	0.851
		0.75	0.038	0.082	0.228	0.465	0.701	0.878
		0.85	0.035	0.073	0.223	0.463	0.691	0.847
		0.95	0.021	0.052	0.178	0.372	0.545	0.629
Normal	$R_n^{(a)}$	-	0.051	0.243	0.741	0.976	0.987	0.992
	$T_n^{(a)}$	-	*0.056*	*0.197*	*0.561*	*0.893*	*0.991*	*0.998*
	$\mathbf{U}_n^{(a)}$	-	0.051	0.174	0.604	0.928	0.994	1.000
	T^2	-	0.050	0.245	0.743	0.981	1.000	1.000
	S_n	0.00	0.041	0.156	0.542	0.884	0.988	0.999
		0.75	0.034	0.182	0.576	0.910	0.968	0.918
		0.85	0.030	0.181	0.572	0.878	0.878	0.743
		0.95	0.017	0.137	0.481	0.664	0.554	0.358
t with 3 d.f.	$R_n^{(a)}$	-	0.050	0.215	0.667	0.923	0.981	0.994
	$T_n^{(a)}$	-	0.054	0.195	0.605	0.893	0.984	0.996
	$\mathbf{U}_n^{(a)}$	-	0.051	0.158	0.493	0.826	0.965	0.995
	T^2	-	*0.041*	*0.160*	*0.578*	*0.878*	*0.969*	*0.991*
	S_n	0.00	0.047	0.171	0.575	0.901	0.987	0.997
		0.75	0.042	0.186	0.617	0.915	0.956	0.903
		0.85	0.037	0.185	0.614	0.870	0.852	0.723
		0.95	0.025	0.150	0.497	0.604	0.509	0.362

was determined from the F distribution table, and for the noninvariant coordinatewise sign test S_n and invariant sign test $T_n^{(\alpha)}$, we used χ^2 approximations (with 2 d.f. and 3 d.f. for $d = 2$ and 3 respectively) for the distributions of the test statistics. In the case of $R_n^{(\alpha)}$, we chose to simulate the permutation distribution of the test statistic as the use of χ^2 approximation to $R_n^{(\alpha)}$ was observed to be not adequate for such a small sample size. The null distribution of the test statistic $\mathbf{U}_n^{(\alpha)}$ is simulated using the method dicussed later in Section 3.4.3. It will be appropriate to note that in the case of invariant sign test, it is possible to simulate the conditional null distribution because of the 'distribution free' nature of the test. However, χ^2 approximation was observed to lead to very good results. The matrix Σ was always estimated using the usual sample variance covariance matrix. Since $R_n^{(\alpha)}$, $T_n^{(\alpha)}$, $\mathbf{U}_n^{(\alpha)}$ and Hotelling's T^2 are all invariant under affine transformations of the data, their powers do not depend on different values of ρ and depend only on the noncentrality parameter $(\theta^T \Sigma^{-1} \theta)^{1/2}$. It is quite clear from the Tables 3.1 and 3.2 that our affine invariant modifications of sign and rank tests have a superior performance over noninvariant procedures for large values of ρ.

3.4.2 *P*-value Computation for Sign and Rank Tests

Let us now consider a real data set and try to investigate the performance of our affine invariant versions of sign and rank tests $T_n^{(\alpha)}$ and $R_n^{(\alpha)}$ applied to it. In a study of cerebral metabolism in epileptic patients Sperling et. al. (1989) measured the metabolic rates of glucose at 10 cortical locations and 6 subcortical locations in the brain by positron emission tomography (PET). Cortical locations were *Frontal, Sensorimotor, Temporal, Parietal* and *Occipital* locations of right and left hemispheres of the brain. Similarly, subcortical locations were *Caudate nucleus, Lenticular nucleus* and *Thalamas* regions of right and left hemispheres. The metabolic rates are measured in mg/100/g/min. We have considered 18 patients forming a normal control group to see whether there is any difference in metabolic rates of right and left hemispheres of the brain. After fixing the transformation matrix $\mathbf{X}(\alpha)$, it is very easy to simulate the null permutation distribution of the invariant rank test statistic, and the P-value of the invariant rank test based on $R_n^{(\alpha)}$ was obtained to be 0.0033 in the case of cortical regions. P-value of invariant sign test based on χ^2 (with 5 d.f.) approximation came out to be 0.046 and those of Hotelling's T^2 and noninvariant sign tests were 0.006 and 0.169 respectively. In the case of subcortical regions, the P-value of invariant rank test turned out to be 0.0264 and that of invariant sign test was 0.014. For the same, classical Hotelling's T^2 test and noninvariant sign test produced P-values that are 0.041 and 0.075 respectively. In both the situations, we see that all the invariant tests conclude that there is a significant difference in metabolic

rates of right and left hemispheres of the brain at 5% level of significance, whereas the noninvariant sign test concludes the difference to be insignificant. This example amply highlights the necessity and importance of using invariant procedures over noninvariant procedures.

Table 3.3: Metabolic rates of glucose at cortical locations in the brain by positron emission tomography

Id	Frontal		Sensorimotor		Temporal		Parietal		Occipital	
	R	L	R	L	R	L	R	L	R	L
1	4.65	4.83	4.88	5.27	4.07	4.13	4.48	4.80	4.36	4.78
2	3.60	3.59	3.28	3.50	3.09	3.25	3.08	3.21	3.00	2.94
3	3.00	3.11	2.77	2.87	2.40	2.48	2.42	2.42	2.42	2.56
4	3.40	3.48	3.41	3.37	3.14	3.12	3.10	3.08	3.36	3.39
5	3.40	3.43	3.29	3.37	3.54	3.47	3.11	3.04	3.49	3.58
6	4.48	4.51	4.84	4.79	4.44	4.33	4.77	4.42	4.53	4.53
7	4.04	3.85	3.67	3.61	3.70	3.19	3.04	2.77	3.47	3.28
8	4.05	3.93	4.11	4.14	3.71	3.67	3.56	3.50	3.88	3.85
9	3.93	4.08	3.96	3.91	3.65	3.91	3.25	3.05	3.41	3.25
10	4.58	4.83	4.59	4.47	3.96	4.28	4.30	4.14	4.50	4.42
11	4.43	4.52	4.36	4.32	3.62	3.44	3.84	3.60	3.48	3.40
12	3.06	3.13	2.96	3.06	2.94	2.89	2.76	2.72	3.15	3.20
13	4.51	4.51	4.20	4.19	4.47	4.41	3.96	3.79	4.51	4.42
14	4.38	4.65	4.41	4.64	4.00	4.06	3.62	4.09	4.57	4.74
15	4.66	4.53	4.24	4.49	4.71	4.36	4.38	4.60	4.88	5.12
16	4.76	4.75	4.67	4.56	4.33	4.32	3.99	3.87	4.30	4.17
17	5.38	5.49	5.18	5.34	4.38	4.36	4.71	4.73	4.56	4.66
18	3.67	3.69	3.81	3.75	4.09	3.90	3.20	3.39	3.68	3.68

3.4.3 *P*-value Computation for Angle Test

At this point we will turn our attention to the affine invariant test introduced and discussed in Section 3.2. It is clear from the proof of Theorem 3.2.3 that once the transformation matrix $\mathbf{X}(\alpha)$ is fixed, the conditional null distribution of $\mathbf{U}_n^{(\alpha)}$ is the same as that of $\sum_{j=1}^{(n-d-1)} \|\mathbf{Y}_j\|^{-1}\mathbf{Y}_j$, where the $\mathbf{Y}_j$'s are i.i.d observations generated from the elliptically symmetric density $\det\{\mathbf{Y}(\alpha)\}f[\mathbf{y}^T\{\mathbf{Y}(\alpha)\}^T\mathbf{Y}(\alpha)\mathbf{y}]$ and $\mathbf{Y}(\alpha) = \Sigma^{-1/2}\mathbf{X}(\alpha)$. Further, elliptic symmetry implies that the distribution of $\|\mathbf{Y}_j\|^{-1}\mathbf{Y}_j$ is uniform on the ellipse, which is completely determined by the matrix $\{\mathbf{Y}(\alpha)\}^T\mathbf{Y}(\alpha) = \{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ and

Table 3.4: Metabolic rates of glucose at subcortical locations in the brain by positron emission tomography

Id	Caudate nucleus		Lenticular nucleus		Thalamas	
	R	L	R	L	R	L
1	4.49	4.73	4.67	4.99	4.58	4.40
2	3.19	3.40	3.09	3.24	2.85	2.81
3	3.02	2.97	3.07	3.12	2.81	2.78
4	3.03	2.83	3.76	4.09	3.41	3.55
5	3.05	3.23	3.97	4.22	3.61	3.54
6	4.55	4.57	5.16	5.34	5.02	4.89
7	3.62	3.48	3.84	3.50	3.85	3.67
8	3.84	4.13	4.18	4.42	3.86	4.15
9	4.08	4.11	4.76	4.56	4.10	4.08
10	4.41	4.55	4.74	4.97	4.64	4.80
11	4.02	3.96	3.83	4.03	3.81	3.71
12	2.73	2.85	3.40	3.52	2.97	2.87
13	4.81	4.91	4.97	5.03	4.69	4.88
14	3.96	4.37	4.05	4.07	4.10	4.15
15	4.44	5.36	6.24	6.20	5.41	5.08
16	4.64	4.64	4.27	4.22	3.91	3.96
17	3.85	4.03	5.16	5.32	4.49	4.64
18	3.18	3.15	4.14	4.36	3.88	3.75

does not depend on f. Hence one can simulate the conditional null distribution of $\mathbf{U}_n^{(\alpha)}$ by taking f to be any specific spherically symmetric density (e.g. the normal density) on $\mathbb{R}^d$. Of course, actual Σ will be unknown in practice, and one can use a consistent affine equivariant estimate $\hat{\Sigma}$ while simulating the null distribution. In the following example, we demonstrate simulation based estimation of the P-value when our test is applied to a real data set.

Example 3.1 : Merchant, Halprin, Hudson, Kilburn, McKenzie, Hurt and Bermazohn (1975) investigated changes in pulmonary functions of twelve workers after they were exposed to cotton dust for six hours. Table 3.5 gives the changes in forced vital capacity (FVC), forced expiratory volume (FEV_3) and closing capacity (CC) for these twelve workers. When Hotelling's T^2 test is applied to this data, the P-value computed using the F distribution turns out to be 0.051. On the other hand the coordinatewise sign test yields a P-value of 0.300 based on a χ^2 approximation with 3 d.f. We estimated the P-value of our test based on a simulation of the conditional null distribution of $\mathbf{U}_n^{(\alpha)}$ using 10,000 Monte Carlo replications, and it turns out to be 0.0721. For simulating the null distribution, we

Table 3.5: Changes in pulmonary functions of twelve workers exposed to cotton dust for six hours.

Subject	FVC	FEV_3	CC
1	−0.11	−0.12	−4.3
2	0.02	0.08	4.4
3	−0.02	0.03	7.5
4	0.07	0.19	−0.3
5	−0.16	−0.36	−5.8
6	−0.42	−0.49	14.5
7	−0.32	−0.48	−1.9
8	−0.35	−0.30	17.3
9	−0.10	−0.04	2.5
10	0.01	−0.02	−5.6
11	−0.01	−0.17	2.2
12	−0.26	−0.30	5.5

have chosen f to be the multivariate spherically symmetric normal density and estimated Σ by the usual variance covariance matrix. Figures in Table 3.5 indicate presence of correlation among the variables, and the scale of the third variable is very different from that of each of the other two. The close agreement between the P-values obtained from two affine invariant tests is noteworthy, and the high P-value produced by the non-invariant coordinatewise sign test is an indication of its failure to detect the deviation from the null hypothesis.

3.4.4 Simulated Powers of Different Two Sample Tests

For two sample problems, we conducted a simulation study similar to that in Section 3.4.1 to judge the performance of our affine invariant two sample rank test based on the statistic $W_{m,n}^{(\alpha)}$. Here we compared the power of invariant rank test with that of invariant two sample median test based on the test statistic $V_{m,n}^{(\alpha)}$, classical Hotelling's T^2 test for two sample location problem and noninvariant median test based on coordinatewise univariate two sample median test statistics $S_{m,n}$. Again the powers of the tests were computed based on 5000 Monte Carlo replications in the case $d = 2$ and 3000 Monte Carlo replications when $d = 3$, and in both the dimensions the sizes of both the samples were taken to be 30 (i.e. $m = n = 30$). The critical value corresponding to the nominal significance level of 5% for Hotelling's T^2 test is obtained from F distribution, and that of invariant and

noninvariant median tests are obtained from χ^2 approximations (with d d.f.). The critical

Table 3.6: Finite sample power of affine invariant rank test and its competitors in the two sample problem for $n = 30$, $d = 2$ and level of significance = 5%.

Distributions	Test	ρ	$(\mu^T\Sigma^{-1}\mu)^{1/2}$					
			0.0	0.3	0.6	0.9	1.2	1.5
Laplace	$W_{m,n}^{(a)}$	-	0.052	0.082	0.226	0.444	0.692	0.864
	$V_{m,n}^{(a)}$	-	0.049	0.085	0.217	0.418	0.651	0.837
	T^2	-	0.047	0.079	0.204	0.423	0.652	0.844
	$S_{m,n}$	0.00	0.041	0.082	0.199	0.410	0.638	0.819
		0.75	0.042	0.080	0.206	0.389	0.633	0.827
		0.85	0.040	0.075	0.181	0.383	0.630	0.829
		0.95	0.039	0.072	0.177	0.384	0.636	0.821
Normal	$W_{m,n}^{(a)}$	-	0.049	0.141	0.467	0.832	0.973	0.999
	$V_{m,n}^{(a)}$	-	0.052	0.109	0.333	0.664	0.894	0.984
	T^2	-	0.051	0.156	0.518	0.874	0.984	1.000
	$S_{m,n}$	0.00	0.046	0.109	0.292	0.661	0.878	0.983
		0.75	0.043	0.106	0.294	0.628	0.889	0.982
		0.85	0.040	0.089	0.286	0.634	0.892	0.982
		0.95	0.039	0.091	0.286	0.629	0.884	0.968
t with 3 d.f.	$W_{m,n}^{(a)}$	-	0.049	0.111	0.317	0.606	0.852	0.956
	$V_{m,n}^{(a)}$	-	0.046	0.095	0.281	0.561	0.804	0.942
	T^2	-	0.046	0.095	0.257	0.508	0.742	0.889
	$S_{m,n}$	0.00	0.050	0.089	0.276	0.503	0.763	0.930
		0.75	0.041	0.081	0.244	0.527	0.788	0.940
		0.85	0.043	0.087	0.243	0.529	0.785	0.944
		0.95	0.041	0.082	0.242	0.524	0.782	0.931

value for invariant rank statistic is obtained by simulating the permutation distribution of the statistic $W_{m,n}^{(a)}$. The results are presented in Tables 3.6 and 3.7 for dimensions $d = 2$ and $d = 3$ respectively for the same three multivariate distributions used earlier. We have selected the optimal transformation matrix $\mathbf{X}(\alpha)$ in the same manner as discussed earlier but this time we have restricted our search in the first sample only (i.e. we have considered only $1 \leq i_k \leq m$ for all $1 \leq k \leq d$) in order to reduce the size of the search problem substantially. From the Tables 3.6 and 3.7, it is clear that invariant two sample rank test has a superior performance over Hotelling's T^2 for non-normal distributions and it is also better than noninvariant procedures for large values of ρ.

Table 3.7: Finite sample power of affine invariant rank test and its competitors in the two sample problem for $n = 30$, $d = 3$ and level of significance = 5%.

Distributions	Test	ρ	$(\mu^T\Sigma^{-1}\mu)^{1/2}$					
			0.0	0.3	0.6	0.9	1.2	1.5
Laplace	$W_{m,n}^{(a)}$	-	0.052	0.074	0.145	0.296	0.506	0.665
	$V_{m,n}^{(a)}$	-	0.052	0.082	0.156	0.265	0.456	0.635
	T^2	-	0.046	0.072	0.141	0.276	0.473	0.664
	$S_{m,n}$	0.00	0.052	0.072	0.145	0.217	0.444	0.632
		0.75	0.053	0.071	0.112	0.221	0.400	0.598
		0.85	0.054	0.078	0.109	0.226	0.408	0.599
		0.95	0.043	0.066	0.108	0.218	0.402	0.588
Normal	$W_{m,n}^{(a)}$	-	0.052	0.129	0.391	0.760	0.955	0.998
	$V_{m,n}^{(a)}$	-	0.054	0.120	0.314	0.633	0.870	0.977
	T^2	-	0.051	0.140	0.447	0.822	0.980	0.999
	$S_{m,n}$	0.00	0.049	0.105	0.313	0.624	0.885	0.959
		0.75	0.045	0.113	0.273	0.595	0.854	0.972
		0.85	0.046	0.113	0.284	0.593	0.856	0.977
		0.95	0.037	0.109	0.276	0.583	0.834	0.936
t with 3 d.f.	$W_{m,n}^{(a)}$	-	0.051	0.098	0.268	0.514	0.772	0.952
	$V_{m,n}^{(a)}$	-	0.054	0.108	0.254	0.504	0.750	0.904
	T^2	-	0.037	0.081	0.228	0.443	0.684	0.853
	$S_{m,n}$	0.00	0.047	0.092	0.211	0.453	0.718	0.891
		0.75	0.047	0.072	0.212	0.486	0.759	0.923
		0.85	0.047	0.073	0.210	0.489	0.764	0.930
		0.95	0.042	0.071	0.211	0.430	0.749	0.893

3.4.5 *P*-value Computation in the Two Sample Case

In order to investigate the performance of the test statistic $W_{m,n}^{(a)}$ when applied to real data, we analyzed a data on the effect of a certain drug on three biochemical compounds found in the brain, which is reported by Morrison (1990, pp. 184–185). 24 mice of the same strain were randomly divided into two equal groups with the second receiving periodic administrations of the drug. Both samples received the same care and diet, and two of the control group mice died of natural causes during the experiment. Assays of the brains of the sacrificed mice revealed the amounts of the compounds in micrograms per gram of brain tissue. We estimated the P-value of the invariant rank statistic by simulating the null permutation distribution of the statistic with 2000 replications, and it came out to

Table 3.8: Effect of a certain drug on three biochemical compound found in the brain of mice

Control			Drug Administered		
1	2	3	1	2	3
1.21	0.61	0.70	1.40	0.50	0.71
0.92	0.43	0.71	1.17	0.39	0.69
0.80	0.35	0.71	1.23	0.44	0.70
0.85	0.48	0.68	1.19	0.37	0.72
0.98	0.42	0.71	1.38	0.42	0.71
1.15	0.52	0.72	1.17	0.45	0.70
1.10	0.50	0.75	1.31	0.41	0.70
1.02	0.53	0.70	1.30	0.47	0.67
1.18	0.45	0.70	1.22	0.29	0.68
1.09	0.40	0.69	1.00	0.30	0.70
			1.12	0.27	0.72
			1.09	0.35	0.73

be 0.0016. P-value of the Hotelling's T^2 based on F distribution was 0.00003, and that of invariant median test and noninvariant median test based on χ^2 approximation were 0.01376 and 0.02894 respectively. Here also there is a noticeable difference in the P-values of the invariant and noninvariant procedures.

3.5 Concluding Remarks

Remark 1: Chaudhuri (1996) and Möttönen and Oja (1995) gave detailed reviews of various notions of multivariate quantiles and ranks. An interesting alternative to our present approach is to use the rank vectors that are associated with spatial median. Affine invariance can still be achieved through data driven transformation as has been done in this Chapter. Such geometric concepts of ranks are very different in nature from coordinatewise ranks considered here. We will discuss in detail these concepts of multivariate ranks in Chapter 5.

Remark 2: Note that once the matrix $\mathbf{X}(\alpha)$ is formed, the computation of the test statistics $T_n^{(\alpha)}$, $R_n^{(\alpha)}$ and $\mathbf{U}_n^{(\alpha)}$ are straightforward. But the selection of optimal α may require a search over $\binom{n}{d+1}$ possible subsets α, and this number grows very fast with n and d. One can reduce the amount of computation involved for searching the optimal

α by stopping whenever the ratio of the arithmetic and geometric means of the eigen values of the matrix $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ is sufficiently close to one. We have observed that this approximation makes the algorithm very fast without making any serious change in the sampling variation of the test statistic or any significant loss of efficiency of the resulting test. In all the real examples that we have considered in Section 3.4, it performed satisfactorily.

Remark 3: It is clear from the definitions that once the transformation matrix $\mathbf{X}(\alpha)$ is fixed, the conditional null distribution of $T_n^{(\alpha)}$ is the same as that of $\sum_{i=1}^{n-d-1} Sign(\boldsymbol{Y}_i)$ and the conditional null distribution of $\mathbf{U}_n^{(\alpha)}$ is the same as that of $\sum_{i=1}^{n-d-1} \|\boldsymbol{Y}_i\|^{-1}\boldsymbol{Y}_i$, where the $\boldsymbol{Y}_i$'s are i.i.d. observations with common density $|\det\{\mathbf{Y}(\alpha)\}|f[y^T\{\mathbf{Y}(\alpha)\}^T\mathbf{Y}(\alpha)y]$ and $\mathbf{Y}(\alpha) = \Sigma^{-1/2}\mathbf{X}(\alpha)$. Further, elliptic symmetry implies that the distributions of these statistics do not depend on f. Hence one can simulate the conditional null distributions of $T_n^{(\alpha)}$ and $\mathbf{U}_n^{(\alpha)}$ by taking f to be any specific spherically symmetric density (e.g. the normal density) on $\mathbb{R}^d$. Of course, actual Σ will be unknown in practice, and one can use a consistent affine equivariant estimate $\hat{\Sigma}$ while simulating the null distribution.

Remark 4: It will be appropriate to note here that in the univariate case, the null distribution of Wilcoxon's signed rank test in the one sample problem and that of rank-sum test in the two sample problem do not depend on the unknown distribution function F of the observations. But this does not hold in the multivariate case. To overcome this difficulty, we consider the basic sign invariance principle which leads to "conditionally distribution-free tests". Under the assumption of elliptic symmetry, the distributions of $\boldsymbol{X}_1$ and $-\boldsymbol{X}_1$ are same, and the joint distribution of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is same as that of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$, where $\boldsymbol{Y}_i$ is $+\boldsymbol{X}_i$ or $-\boldsymbol{X}_i$, $1 \leq i \leq n$. Given the observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, we consider the permutation distribution over 2^n possible points. Puri and Sen (1971) showed that the permutation tests have the same size as the unconditional test. We have simulated the permutation distributions of our invariant signed rank and rank-sum test statistics and based on them we have computed the powers of the tests.

Chapter 4

Multivariate Linear Models

4.1 Least Absolute Deviations and Rank Regression

Consider a linear regression set-up with a k-dimensional regressor $\boldsymbol{x}$ and a univariate response y satisfying the linear model $y = \beta^T \boldsymbol{x} + e$, where our objective is to estimate and make inference about the k-dimensional parameter vector β based on a set of independent observations $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$. In this set-up, the method of least absolute deviations (*LAD*) and the method of least squares (*LS*) have competed with each other for more than two hundred years. LAD estimation technique is known to have greater antiquity than least squares method (see e.g. Bloomfield and Steiger 1983). Legendre published his "Principle of Least Squares" in 1805. But nearly half a century earlier, sometime between 1755 and 1757, R.J. Boscovitch discussed an interesting criterion for fitting a line to $n > 2$ points in the plane (see Eisenhart 1961), which is nothing but choosing a line by minimizing the sum of absolute deviations between the observed and fitted y-values among all lines constrained to pass through the mean of the data points. In 1760, he outlined a simple geometric algorithm to find a solution to this constrained minimization problem, which was algebraically formalized by Laplace in 1789. For a long period, no good algorithm for computing LAD estimates in a general set-up was available even when $k = 2$. *LS estimates certainly did not* have this drawback as they can be expressed as simple and closed form solutions to certain systems of linear equations, and this has greatly contributed towards overwhelming popularity of *LS over LAD among* practitioners from the very inception of LS. Another serious difficulty with LAD estimation was that the distributional properties of the resulting estimates were not easy to work out analytically, whereas those of LS estimates were well known and easy to use for the purpose of making statistical inference. Bassett and Koenker (1978) investigated LAD estimates in linear models and proved several interesting results related to it. Since then

a vast amount of literature has evolved extending the notion of LAD estimation in various directions in the linear regression set-up with a univariate response. Koenker and Bassett (1978) proposed and investigated quantile regression in linear models. Ruppert and Carroll (1980) considered two methods of defining regression analogs of the trimmed mean. The first one was originally suggested by Koenker and Bassett (1978) and uses their concept of regression quantiles. Its asymptotic behavior is completely analogous to that of a trimmed mean. The second method uses residuals from a preliminary estimator, and its asymptotic behavior heavily depends on that preliminary estimate. Welsh (1987) proposed another analog of trimmed mean using von Mises functional approach, and he established asymptotic and robustness properties of the proposed estimate, which are equivalent to those of the estimate proposed by Koenker and Bassett (1978). It is well known now that the LAD regression problem can be formulated as a linear programming problem, and as a result, several good algorithms are available for computing the LAD estimates (Armstrong and Kung 1978, Barrodale and Roberts 1973, Bloomfield and Steiger 1983, Koenker and d'Orey 1987, Narula and Wellington 1977, Wesolowsky 1981). For estimating the parameters of a structural equation in a simultaneous equation model, Amemiya (1982) extended LAD estimators to two-stage least absolute deviation estimators and established the strong consistency and asymptotic normality of the estimates. Subsequently McKean and Schrader (1987), Schrader and McKean (1987), Bai, Rao and Yin (1990) showed that the statistical inference procedures based on LAD estimates are quite similar to the classical analysis of variance based on least squares. Here the reduction in sum of squares is replaced by the reduction in sum of absolute errors, which leads to summarization of the analysis in the form of 'LAD analysis of variance table'. Strong consistency of LAD estimates and their Bahadur-type representations have been established by Babu (1989) and also discussed as a special case to a very general result obtained by Neimiro (1992).

In linear models with univariate response, rank regression techniques have been proposed and extensively studied as alternatives to traditional least squares regression by several statisticians [see e.g. Lehmann (1963a, 1963b, 1964), Adichi (1967, 1978), Koul (1969, 1970), Puri and Sen (1969, 1973), Jureckova (1971, 1973), Jaeckel (1972), Hettmansperger and McKean (1977, 1978, 1983)]. These authors explored various extensions of rank based methods, which were originally developed for nonparametric inference in one and two sample univariate location problems, into very general linear models including standard ANOVA models. A primary motivation behind considering rank regression is the lack of robustness in least squares regression, which is known to have very poor performance when the random error e in the linear model happens to follow non-Gaussian distributions especially those with heavy tails. Higher statistical efficiency of rank based nonparametric procedures compared to the inference based on sample means in one and two sample lo-

cation problems involving univariate non-Gaussian data is known to extend for parameter estimates and related inference based on rank regression in linear models with univariate response. An excellent review of various rank based statistical methods for linear models with real valued response can be found in Draper (1988) and in fascinating comments and discussion that Draper's expository article was successful in generating from leading experts in robust regression in linear models.

But so far all the work documented in the literature is restricted to essentially univariate response y. Almost nothing exists in the literature beyond least squares methods when we have a d-dimensional ($d > 1$) response vector y, and the problem is to estimate the $k \times d$ dimensional matrix of parameters β in the multiresponse linear regression model $y = \beta^T x + e$. To motivate the need for considering such a multiresponse linear regression, let us consider the following example.

Table 4.1: Systolic and diastolic blood pressure distribution with age of 40 Marwari females residing in Calcutta

Serial no.	Age	Systolic pressure	Diastolic pressure	Serial no.	Age	Systolic pressure	Diastolic pressure
1	52	130	80	21	26	130	84
2	21	120	88	22	76	160	90
3	60	180	100	23	37	110	80
4	38	110	90	24	48	130	90
5	19	100	70	25	40	160	112
6	50	170	100	26	36	150	90
7	32	130	84	27	39	140	100
8	41	120	80	28	38	110	74
9	36	140	84	29	16	110	70
10	57	170	106	30	48	130	100
11	52	110	80	31	22	120	80
12	19	120	80	32	30	110	70
13	17	110	70	33	19	120	80
14	16	120	80	34	39	124	84
15	67	160	90	35	38	130	94
16	42	130	90	36	45	120	84
17	44	140	90	37	22	130	80
18	56	170	100	38	20	120	86
19	32	150	94	39	18	120	80
20	21	140	94	40	31	112	80

Example 4.1: Biological Sciences Division of Indian Statistical Institute, Calcutta

collected data on blood pressures of 40 Marwari females residing at Burrabazar area of Calcutta (see Table 4.1). It is well-known to physiologists that the arterial pressures tend to increase with the age of a person. Several empirical studies have been made in this context, and it has been observed that this relationship depends on various environmental factors as well as the ethnic status of a person. In other words, there is no common relationship that works for all human beings, and it is different for different groups of people. Nevertheless it is accepted by all physiologists that the age is an important factor in deciding what should be the normal blood pressures of a person.

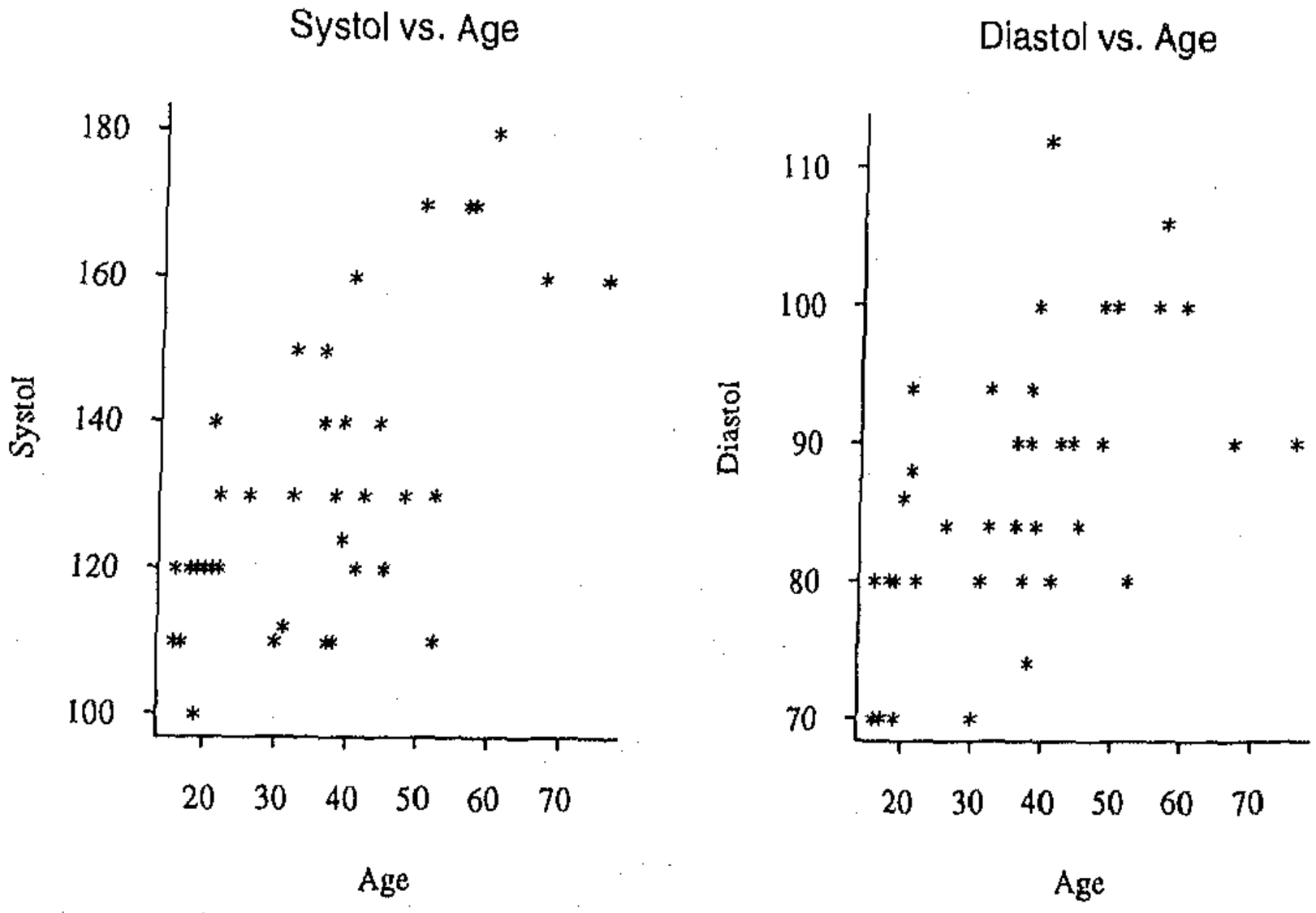


Figure 4.1: The systolic and diastolic blood pressures are plotted against the age of 40 women. The plot shows high variability in the scattered data with some possible outliers.

We are interested here in finding an empirical relationship of systolic and diastolic blood pressures with the age of a normal Marwari woman residing at Calcutta. Now in Figure 4.1, which shows the scatter plots of systolic and diastolic blood pressures against age, there are some outlying observations and the spread of the data is very high. It is well-known that unlike LAD estimates, the LS estimates of the regression parameters are highly sensitive to outlying data points (e.g. those corresponding to very high or low blood pressures). One can argue that very high or low pressure cases should not have any undue

influence on an empirically developed relationship of the blood pressures with the age of a normal female. This is one of the primary reasons for using an appropriate extension of the LAD method that will be suitable for this multiresponse regression problem.

Rao (1988) addressed the problem of generalizing LAD estimation in multiresponse linear regression set-up and suggested the use of univariate least absolute deviation regression for each coordinate of the response vector. He has shown that under simple conditions that estimate is asymptotically normal, but the problem with his estimation technique is that it does not take into account the inter-dependence of the coordinates of the response vector, and it may not be always wise to ignore correlations that exist among different response variables. Another approach to generalise LAD estimation in the multivariate response problem is due to Bai, Chen, Miao and Rao (1990), who extended the notion of spatial median (cf. Brown 1983, Chaudhuri 1992a, Haldane 1948) in the regression set-up, and obtained their estimate by minimizing $\sum_{i=1}^{n} \|y_i - \beta^T x_i\|$ w.r.t. β (here $\|\cdot\|$ denotes the usual Euclidean norm). It is easy to observe that while in the univariate case, this leads to estimates that are equivariant under the scale transformation of the response variable, in the case of multivariate response, the estimated parameter matrix will not be equivariant under arbitrary nonsingular linear transformations of the response vector. It is known that for multivariate data with correlated variables, spatial median may have poor statistical efficiency compared to affine invariant sample mean vector [see Brown (1983), Chaudhuri (1992a), see also Chapter 2]. Further, the lack of scale equivariance makes spatial median as well as its generalization in linear models practically useless when different real valued components of the response vector y are measured in different scales or when the response variables have different degrees of statistical variation. In another generalization, Koenker and Portnoy (1990) suggested M-estimation in the multi-response linear regression model. Though their generalization has some nice properties, it fails to be affine equivariant and they have discussed the lack of affine equivariance of their estimates and related matters in some details. Davis and McKean (1993) have extensively studied the coordinatewise extension of rank regression from univariate to multivariate response problems. These authors have derived some interesting statistical properties of their proposed estimates and tests and reported some results on numerical performance of the procedures. This coordinatewise extension of rank regression too fails to take into account the inter-dependence among the real valued components of the response vector, and in practice it may not be appropriate to ignore the correlation present among different response variables. Procedures that lack affine equivariance/invariance are known to have very poor statistical performance in the presence of substantial correlation among the components of the response vector. [see Chapters 2 and 3].

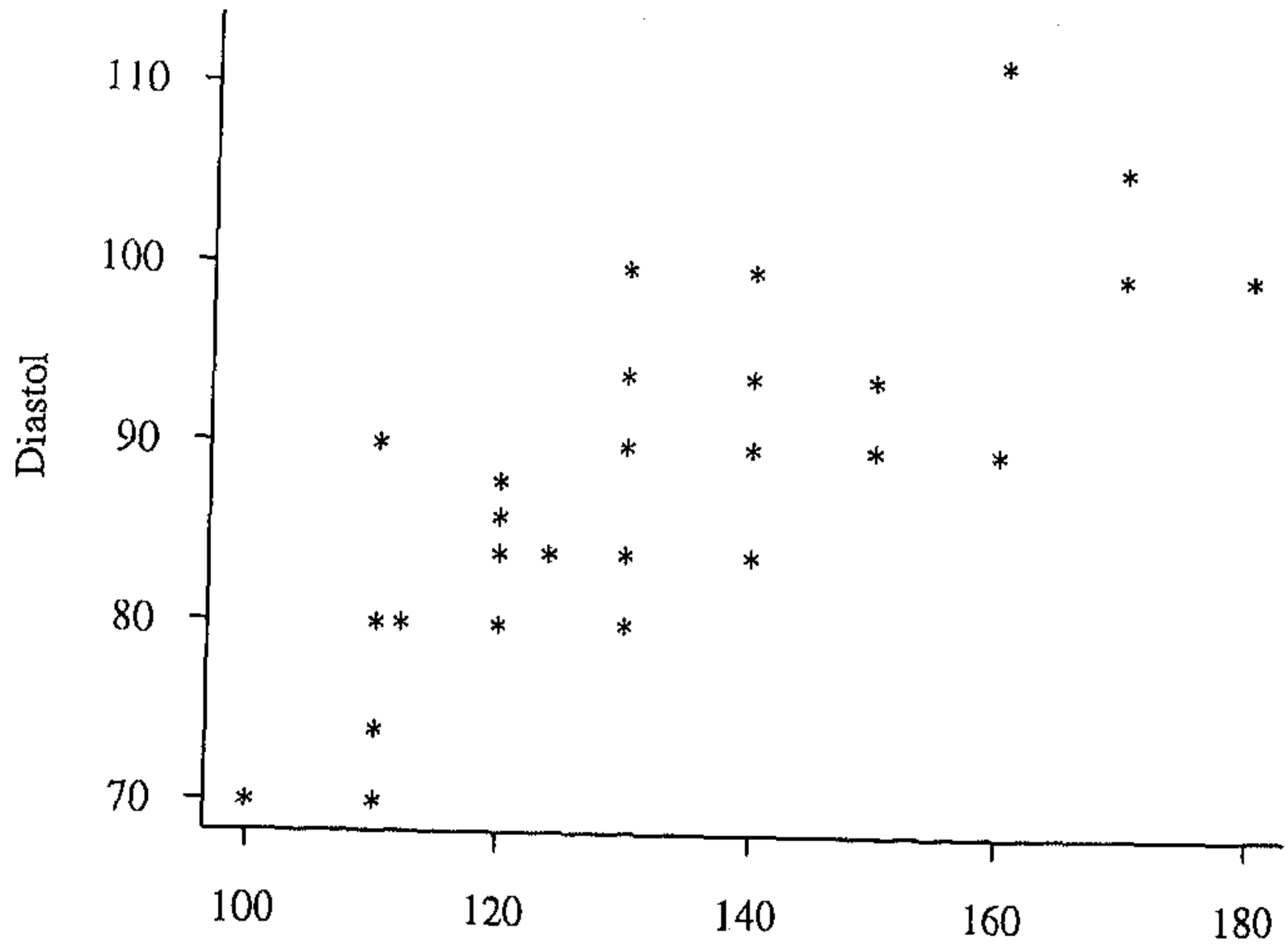


Figure 4.2: The scatter plot showing the distribution of systolic blood pressures against diastolic blood pressures.

Let us consider the blood pressures data discussed in Example 4.1. It is fairly obvious from Figure 4.2 that there exists high correlation between the systolic and the diastolic pressures of a person as one would expect, and instead of using the least absolute deviation regression for each of the two pressure measurements separately, we need to use some affine equivariant approach.

Example 4.2: Srinivasan (1995) compiled fertility and mortality figures from the official publications of the Registrar General of India and figures on female literacy rates from Decennial Census reports. The data-set contains total fertility rates (TFR), infant mortality rates (IMR) and female literacy rates (FLR) for the years 1971, 1981 and 1991

for sixteen most populated states of India. (See Table 4.2). Our interest here is in exploring

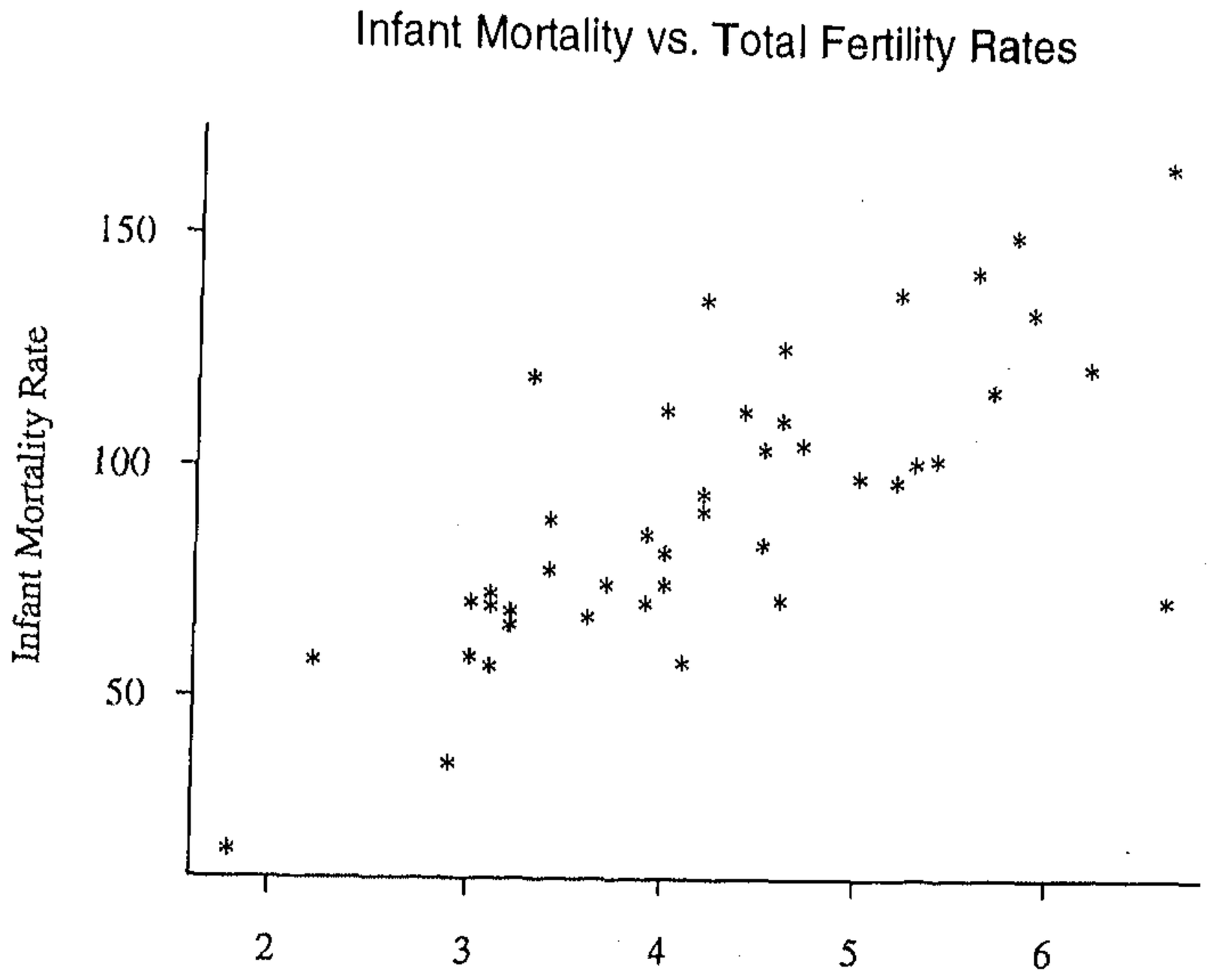


Figure 4.3: The total fertility rate (TFR) is plotted against infant mortality rate (IMR) showing high correlation among themselves.

relationships of TFR and IMR with FLR over different years and regions. As the data-set contains only 48 observations, we clubbed the states into 4 regions namely North, South, East and West instead of looking at them separately. It is well-known from socio-demographic studies that TFR and IMR are highly correlated (see Figure 4.3). So in this situation too the asymptotic efficiency of the non-equivariant coordinate-wise LAD estimates is likely to be very poor, and one needs to use affine equivariant estimates of regression parameters that will take into account the inter-dependence between TFR and IMR.

In this Chapter, we propose and investigate in detail a technique for estimating pa-

Table 4.2: Fertility, infant mortality and female literacy rates for different states of India during 1971-1991

States	Total Fertility Rates			Infant Mortality Rates			Female Literacy Rates		
	1971	1981	1991	1971	1981	1991	1971	1981	1991
Andhra Pradesh	4.7	3.9	3.0	106	86	71	19.2	24.2	33.7
Assam	5.2	4.1	3.4	139	104	78	22.8	–	43.7
Bihar	5.0	5.7	4.6	–	118	72	10.7	16.5	23.1
Gujarat	5.6	4.4	3.2	144	113	69	30.3	38.5	48.5
Haryana	6.6	5.0	3.9	72	99	71	18.6	26.9	40.9
Himachal Pradesh	4.4	4.0	3.1	113	75	70	24.7	37.7	52.5
Karnataka	4.2	3.6	3.1	95	68	73	25.7	33.2	44.3
Kerala	4.1	2.9	1.8	58	36	17	65.4	75.7	86.9
Madhya Pradesh	5.9	5.2	4.6	135	139	111	13.7	19.0	28.4
Maharashtra	4.5	3.7	3.0	105	75	59	32.4	41.0	50.5
Orissa	4.6	4.2	3.3	127	137	120	17.0	25.1	34.4
Punjab	5.3	4.0	3.1	102	82	57	31.3	39.6	49.7
Rajasthan	6.2	5.4	4.5	123	103	84	10.5	14.0	20.8
Tamilnadu	4.0	3.4	2.2	113	89	58	32.3	40.4	52.3
Uttar Pradesh	6.6	5.8	5.2	167	152	98	13.0	17.2	26.0
West Bengal	–	4.2	3.2	–	91	66	27.8	36.1	47.2

rameters in linear models with multivariate response, which will lead to estimates that are equivariant under nonsingular linear transformations of the response vector. First observe that the LAD regression problem is a median regression problem in the sense that the conditional median of the response y given the regressors x is being estimated, whereas LS regression problem is a mean regression problem, where the conditional mean of y given x is estimated. So, in order to solve the multiresponse LAD linear regression problem, one needs to define a proper analog of the median in multidimension. In the preceding Chapters we have already discused several definitions of multivariate median available in the literature. The vector of coordinate-wise median lacks the property of equivariance even under orthogonal transformations, and its regression analog, which is the coordinate-wise LAD estimates, has the same drawback. Spatial median is equivariant under orthogonal transformations but not under arbitrary affine transformations, and the same is true for the regression estimates proposed by Bai et al.(1990). There are several definitions of multivariate medians that are affine equivariant in nature (see Liu 1990, Oja 1983, Tukey 1975), and each of them leads to a regression analog, which will be equivariant under nonsingular linear transformations of the response. However, none of these regression analogs has been considered in the existing literature, and all of them are computationally

so intensive that having estimates of regression parameters may turn out to be virtually impossible in practice with the available computing resources even when both the sample size and the dimension of the parameter space are only moderately large. We consider here the regression analog of the transformation retransformation coordinatewise median defined in Chapter 2. We will demonstrate that the proposed estimate outperforms the matrix of coordinate-wise LAD estimates when the real valued components of the response vector are correlated. The procedure suggested in this paper is easy to compute, and we provide a convenient algorithm, which enables one to compute parameter estimates as well as to invoke resampling strategies such as the bootstrap to estimate finite sample variance covariance matrix of the estimates.

In Section 4.2, we pose the multiresponse linear regression problem in detail with necessary assumptions and describe the methodology as well as the computation of the estimate of the parameter matrix. We discuss a simple algorithm called TREMMER to compute our proposed transformation retransformation LAD estimates. Then we demonstrate with two real examples, the performance of the procedure. In Section 4.3, we establish some important asymptotic results about the proposed estimate and demonstrate some optimal properties of the transformation retransformation median regression estimates.

In Section 4.4 that follows, we will describe how one can appropriately modify TREMMER to come up with affine equivariant rank regression procedures for multi-response linear models. Such a modification inherits the nice statistical properties of TREMMER, and in the special case of regression based on Wilcoxon's rank scores or equivalently the linear regression analogs of Hodges-Lehmann type estimates [see e.g. Chaudhuri (1992b)], this modification takes a simplified and elegant form that makes the implementation of the methodology as well as investigation into its statistical properties quite convenient. In Section 4.5, we will present some results based on numerical studies that were undertaken to investigate the performance of the proposed methodology. We will discuss results from small sample simulation experiments that yield strong evidence for superior performance of transformation retransformation rank regression estimates in multi-response linear model problems when compared with traditional least squares and coordinatewise least absolute deviations estimates if the residuals in the linear model have elliptic non-Gaussian distributions with heavy tails. We will also report analysis of real data sets in an attempt to demonstrate the implementation of the methodology in real data and how it outperforms some competing non-equivariant procedures.

4.2 Description and Computation of TREMMER Estimates

Consider the following multiresponse linear model :

$$y_i = \beta^T x_i + e_i, \qquad i = 1, \ldots, n \tag{4.1}$$

where the y_i's are $d \times 1$ response vectors, the x_i's are $k \times 1$ dimensional vectors of explanatory variables, the e_i's are d-dimensional error vectors, and β is a $k \times d$ matrix of parameters. We assume that the e_i's are independent and identically distributed with a common probability distribution on $\mathbb{R}^d$. Before defining the transformation retransformation strategy, let us observe a simple geometrical fact about any given affine transformation of a set of multivariate responses. For a nonsingular $d \times d$ matrix $\mathbf{A}$, the transformation that maps y_i into $\mathbf{A}y_i$ for $1 \leq i \leq n$ essentially expresses the original linear model in terms of a new coordinate system determined by $\mathbf{A}$ and depending on whether $\mathbf{A}$ is an orthogonal matrix or not, this new coordinate system may or may not be an orthonormal system. The fundamental idea that lies at the root of the data based transformation retransformation is to form an appropriate 'data driven coordinate system' [see Chapter 2] and to express the linear model in terms of that coordinate system first. This is equivalent to making an affine transformation of the error vectors. Then one computes parameter estimates based on transformed response vectors. Finally the estimates of regression parameters are retransformed to express everything back in terms of the original coordinate system. Now, in order to form a 'data driven coordinate system', we need d points in $\mathbb{R}^d$ and the lines joining the origin to these d points will form various coordinate axes. In order to get a valid coordinate system, these d points must satisfy some nonsingularity condition.

Let us define

$$A_n = \{a : a \subset \{1, 2, \ldots, n\} \text{ and } \#\{i : i \in a\} = k\}$$

$$B_n = \{b : b \subset \{1, 2, \ldots, n\} \text{ and } \#\{i : i \in b\} = d\}$$

and

$$S_n = \{\alpha = a \cup b : a \in A_n \quad b \in B_n, \quad a \cap b = \phi\}.$$

Note that S_n is the set of all subsets of $k+d$ indices from the set $\{1, 2, \ldots, n\}$. For a fixed $\alpha = \{i_1, \ldots, i_k, j_1, \ldots, j_d\} \in S_n$, let $\mathbf{W}(\alpha)$ be the $k \times k$ matrix whose columns are the vectors $x_{i_1}, \ldots, x_{i_k}$, and $\mathbf{Z}(\alpha)$ be the $d \times k$ matrix whose columns are the vectors $y_{i_1}, \ldots y_{i_k}$. We will assume that $\mathbf{W}(\alpha)$ is invertible and define $\mathbf{E}(\alpha)$ to be the $d \times d$ matrix consisting of the columns $y_{j_1} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}x_{j_1}, \ldots, y_{j_d} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}x_{j_d}$. If the error vectors e_i's happen to be i.i.d random vectors with a common probability distribution, which is

absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, the matrix $\mathbf{E}(\alpha)$ will be invertible with probability one. Then define transformed response vectors $z_j^{(\alpha)} = \{\mathbf{E}(\alpha)\}^{-1} y_j$ for $1 \leq j \leq n$ with $j \notin \alpha$. Let $\hat{\Gamma}_n^{(\alpha)}$ be the matrix of parameter estimates obtained by regressing each coordinate of the d-dimensional transformed vectors $z_i^{(\alpha)}$'s separately on the x_i's for $1 \leq i \leq n$ and $i \notin \alpha$ using the LAD method. Then define the transformation retransformation estimate of the parameter matrix as $\hat{\beta}_n^{(\alpha)} = \hat{\Gamma}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^T$. Note that $\hat{\beta}_n^{(\alpha)}$ is obtained by retransforming the earlier $\hat{\Gamma}_n^{(\alpha)}$ by the linear transformation $\mathbf{E}(\alpha)$. The following Theorem asserts equivariance of $\hat{\beta}_n^{(\alpha)}$ under nonsingular transformations of the response vector and the regression equivariance of it.

Theorem 4.2.1 *For a fixed $\alpha \in S_n$, let $\hat{\beta}_n^{(\alpha)}$ be the estimated matrix of parameters based on the data-points $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ as described above.*

(i) Suppose that $\mathbf{A}$ is a fixed $d \times d$ nonsingular matrix. Then the transformation-retransformation estimate computed from $(\mathbf{A}y_1, x_1), \ldots, (\mathbf{A}y_n, x_n)$ in the same way as above (using the same index set α) will be $\hat{\beta}_n^{(\alpha)}\mathbf{A}^T$.

(ii) Suppose that the response vectors y_i's are transformed to $y_i - \mathbf{G}^T x_i$ for $i = 1, \ldots, n$ where $\mathbf{G}$ is a fixed $k \times d$ matrix. Then the transformation retransformation estimate will be transformed to $\hat{\beta}_n^{(\alpha)} - \mathbf{G}$.

Proof: (i) First observe that in view of the way the matrix $\mathbf{Z}(\alpha)$ has been constructed, if the y_i's are transformed to $\mathbf{A}y_i$'s, $\mathbf{Z}(\alpha)$ will be transformed to $\mathbf{A}\mathbf{Z}(\alpha)$. In turn the transformation matrix $\mathbf{E}(\alpha)$ is transformed to $\mathbf{A}\mathbf{E}(\alpha)$. Also, note that the $z_i^{(\alpha)}$'s remain invariant under a nonsingular linear transformation of the y_i's. Hence, the estimated matrix of regression parameters $\hat{\Gamma}_n^{(\alpha)}$ obtained by regressing each coordinate of the $z_i^{(\alpha)}$'s on the x_i's using LAD separately is invariant under that transformation. Consequently, $\hat{\beta}_n^{(\alpha)}$, which was originally defined as $\hat{\Gamma}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^T$, will be transformed to $\hat{\beta}_n^{(\alpha)}\mathbf{A}^T$.

(ii) Observe that if the y_i's are transformed to $y_i - \mathbf{G}^T x_i$'s, the matrix $\mathbf{Z}(\alpha)$ will be transformed to $\mathbf{Z}(\alpha) - \mathbf{G}^T\mathbf{W}(\alpha)$. In turn the transformation matrix $\mathbf{E}(\alpha)$ remains invariant. Also, note that the $z_i^{(\alpha)}$'s are transformed to $z_i^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}^T x_i$. Hence, the estimated matrix of regression parameters $\Gamma_n^{(\alpha)}$ obtained by regressing each coordinate of the $z_i^{(\alpha)}$'s on the x_i's using LAD will be transformed to $\Gamma_n^{(\alpha)} - \mathbf{G}\{[\mathbf{E}(\alpha)]^T\}^{-1}$ in view of the regression equivariance property of LAD. Consequently, $\hat{\beta}_n^{(\alpha)}$ will be transformed to $\hat{\beta}_n^{(\alpha)} - \mathbf{G}$. □

From the definition of the transformation retransformation estimate of the matrix of parameters and from the above Theorem, we make the following simple observations :

Observation 1 : If $k = 1$ and the regressors $x_i = 1$, for $1 \leq i \leq n$, then our problem

reduces to estimation of the multivariate median of the observations $y_1, y_2, \ldots, y_n$, and our estimation procedure leads to transformation retransformation multivariate coordinatewise median introduced in Chapter 2.

Observation 2 : The transformation retransformation estimate of the parameter matrix is obtained as the minimizer of $\sum_{i \in \alpha} |\{\mathbf{E}(\alpha)\}^{-1}(y_i - \beta^T x_i)|$ with respect to $\beta \in \mathbb{R}^{k \times d}$, where $|\cdot|$ denotes the l_1 norm in $\mathbb{R}^d$. This implies that the estimate $\hat{\beta}_n^{(\alpha)}$ is equivariant under linear reparametrization of the design points x_i's. In other words, if we transform our regressor vectors x_i's to $\mathbf{B}x_i$'s for $1 \leq i \leq n$, where $\mathbf{B}$ is a $k \times k$ nonsingular matrix, our estimate is transformed to $(\mathbf{B}^T)^{-1}\hat{\beta}_n^{(\alpha)}$.

Observation 3 : Consider the multiresponse linear model with an intercept term $\gamma \in \mathbb{R}^d$, i.e.

$$y_i = \gamma + \beta^T x_i + e_i, \qquad i = 1, \ldots, n.$$

If the y_i's are transformed to $y_i + \mathbf{b}$, where $\mathbf{b}$ is a d-dimensional vector, then by *(ii)* of Theorem 4.2.1 our transformation retransformation estimate $\hat{\gamma}_n^{(\alpha)}$ will be transformed to $\hat{\gamma}_n^{(\alpha)} + \mathbf{b}$ and $\hat{\beta}_n^{(\alpha)}$ will remain unchanged.

4.2.1 Asymptotic Normality and Selection of α

Clearly, for different choices of the subset of indices α, we have different estimates of the parameter matrix β. So the natural question that arises at this stage is which subset of indices α we should use. Our approach for selecting the subset α is based on the minimization of the generalized variance (Wilks 1932) of the estimate $\hat{\beta}_n^{(\alpha)}$, which is defined as the determinant of the variance covariance matrix of the estimate. Recall that this determinant is proportional to the volume of the concentration ellipsoid associated with the sampling distribution of the estimate. If we assume that the underlying common probability distribution of the error vector e is elliptically symmetric with a density of the form $\{\det(\Sigma)\}^{-1/2} f(e^T \Sigma^{-1} e)$ where Σ is a $d \times d$ positive definite matrix, and $f(e^T e)$ is a spherically symmetric density on $\mathbb{R}^d$, we have a nice simple form for the asymptotic generalized variance of the estimate $\hat{\beta}_n^{(\alpha)}$ for a given α as given in the following Theorem. Let us write $\{\Sigma^{-1/2}\mathbf{E}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$, where $\mathbf{R}(\alpha)$ is a diagonal matrix with positive diagonal entries, and $\mathbf{J}(\alpha)$ is a matrix whose rows are of unit length. The following Theorem gives the asymptotic distribution of the estimated regression parameter matrix $\hat{\beta}_n^{(\alpha)}$.

Theorem 4.2.2 *Fix $\alpha \in S_n$. Assume that the density f is such that any univariate marginal g of the spherically symmetric density $f(\mathbf{e}^T\mathbf{e})$ is differentiable and posi-*

tive at zero, and $\max_{i \notin \alpha} \boldsymbol{x}_i^T \{\sum_{j \notin \alpha} \boldsymbol{x}_j \boldsymbol{x}_j^T\}^{-1} \boldsymbol{x}_i$ *converges to zero as* n *tends to infinity. Then, as* n *tends to infinity, the conditional distribution of* $\{\sum_{j \notin \alpha} \boldsymbol{x}_j \boldsymbol{x}_j^T\}^{1/2} (\hat{\boldsymbol{\beta}}_n^{(\alpha)} - \boldsymbol{\beta})$ *given the* $\boldsymbol{e}_i$*'s with* $i \in \alpha$ *converges weakly to a multivariate normal distribution with zero mean and* $c\Sigma^{1/2}\mathbf{V}(\alpha)\Sigma^{1/2} \otimes \mathbf{I}_k$ *as the dispersion matrix. Here* $c = \{2g(0)\}^{-2}$, $\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^T\}^{-1}$, *and* $\mathbf{D}(\alpha)$ *is the* $d \times d$ *matrix whose* (i,j)*-th element is* $(2/\pi)\sin^{-1}\gamma_{ij}$, γ_{ij} *being the inner product of the* i*-th and the* j*-th row of* $\mathbf{J}(\alpha)$. *We denote by* $\otimes$ *the usual Kronecker product, and* $\mathbf{I}_k$ *is the identity matrix of dimension* $k \times k$.

Proof: In view of the equivariance of the regression estimates $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ under nonsingular linear transformations of the $\boldsymbol{y}_i$'s, it is sufficient to prove the Theorem in the special case when Σ is the $d \times d$ identity matrix. Define $\boldsymbol{e}_i^* = \{\mathbf{E}(\alpha)\}^{-1}\boldsymbol{e}_i$ for $1 \leq i \leq n$ and $i \notin \alpha$ to be the transformed error vectors. Then, given the $\boldsymbol{e}_i$'s for which $i \in \alpha$, we have the transformed model as,

$$\boldsymbol{z}_i^{(\alpha)} = \Gamma^T \boldsymbol{x}_i + \boldsymbol{e}_i^*, \qquad i \notin \alpha. \tag{4.2}$$

Under the assumption that $\max_{i \notin \alpha} \boldsymbol{x}_i^T \{\sum_{j \notin \alpha} \boldsymbol{x}_j \boldsymbol{x}_j^T\}^{-1} \boldsymbol{x}_i$ converges to zero as n tends to infinity, we have the following representation (see Babu 1989)

$$2g_i(0)\{\sum_{j \notin \alpha} \boldsymbol{x}_j \boldsymbol{x}_j^T\}^{1/2}(\hat{\Gamma}_{in}^{(\alpha)} - \Gamma_i) = \sum_{j \notin \alpha}\{\sum_{l \notin \alpha} \boldsymbol{x}_l \boldsymbol{x}_l^T\}^{-1/2} \boldsymbol{x}_j \operatorname{sign}(U_{ji}) + \mathbf{R}_n,$$

where U_{ji} is the i-th component of $\boldsymbol{e}_j^*$, g_i is the i-th marginal density of the distribution of $\boldsymbol{e}_j^*$ and $\hat{\Gamma}_{in}^{(\alpha)}$ and Γ_i are the i-th columns of $\hat{\Gamma}_n^{(\alpha)}$ and Γ respectively. Here $\mathbf{R}_n$ converges in probability to zero. By the assumption on the $\boldsymbol{x}_i$'s stated in the Theorem, the Lindeberg condition for CLT is satisfied for the first term on the right hand side, and hence we have the asymptotic normality of the estimated regression parameter matrix given the $\boldsymbol{e}_i$'s for which $i \in \alpha$. Note that we have not used the elliptic symmetry of the error distribution. In other words, asymptotic normality holds in a large class of probability distributions.

Now, under the assumption of elliptic symmetry of the error distribution as stated in the Theorem, $\boldsymbol{e}_i^*$'s with $i \notin \alpha$ are conditionally i.i.d random vectors with common density $|\det\{\mathbf{E}(\alpha)\}| f\{\boldsymbol{e}^T[\mathbf{E}(\alpha)]^T[\mathbf{E}(\alpha)]\boldsymbol{e}\}$. Let $r_1, \ldots, r_d$ be the diagonal entries of $\mathbf{R}(\alpha)$. In view of the above representation, the conditional distribution of $\{\sum_{j \notin \alpha} \boldsymbol{x}_j \boldsymbol{x}_j^T\}^{1/2}(\hat{\Gamma}_n^{(\alpha)} - \Gamma)$ will converge weakly to a $k \times d$-variate normal distribution with zero mean, and limiting dispersion matrix $\mathbf{S}(\alpha) \otimes \mathbf{I}_k$. Here the matrix $\mathbf{S}(\alpha)$ is such that its i-th diagonal entry is cr_i^2, and for $i \neq j$, its (i,j)-th element is $4cr_ir_j\{Pr(U_{lj} < 0 \text{ and } U_{li} < 0) - 1/4\}$. U_{li} and U_{lj} are the i-th and the j-th components of $\boldsymbol{e}_l^*$ respectively. Note that we are using the fact that for a d-dimensional random vector $\boldsymbol{z}$ with a spherically symmetric distribution, the distribution of the random variable $\boldsymbol{a}^T\boldsymbol{z}$ is the same for any $\boldsymbol{a} \in \mathbb{R}^d$ such that $\boldsymbol{a}^T\boldsymbol{a} = 1$.

Also, since the conditional distribution of $\boldsymbol{e}_i^*$ is elliptically symmetric around the origin in $\mathbb{R}^d$, $Pr\{U_{lj} < 0 \text{ and } U_{li} < 0\}$ does not depend on the density f. Recall that the rows of $\mathbf{J}(\alpha)$ are of unit length obtained by normalizing the rows of $\{\mathbf{E}(\alpha)\}^{-1}$. We now have the following by some routine analytic computation,

$$Pr(U_{lj} < 0 \text{ and } U_{li} < 0) = 1/4 + (1/2\pi)\sin^{-1}\gamma_{ij}.$$

So, the matrix $\mathbf{S}(\alpha)$ is nothing but $c\{\mathbf{R}(\alpha)\}\{\mathbf{D}(\alpha)\}\{\mathbf{R}(\alpha)\}$. Next recall that

$$\hat{\boldsymbol{\beta}}_n^{(\alpha)} = \hat{\Gamma}_n^{(\alpha)}\{\mathbf{R}(\alpha)\}^{-1}\{[\mathbf{J}(\alpha)]^T\}^{-1}.$$

The proof of the Theorem is now complete by straightforward algebra. □

It follows from the preceding Theorem that $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ is a $n^{1/2}$-consistent estimate of $\boldsymbol{\beta}$, and its conditional asymptotic generalized variance is

$$[c^d\{\det(\Sigma)\}\det\{\mathbf{V}(\alpha)\}]^k[\det\{\sum_{i\notin\alpha}\boldsymbol{x}_i\boldsymbol{x}_i^T\}]^{-d}$$

Corollary 4.2.3 *Suppose that the conditions on the distribution of the error vector e stated in Theorem 4.2.2 are satisfied, and assume that* $n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$ *converges to a positive definite matrix* $\mathbf{Q}$ *as* n *tends to infinity. Then the conditional distribution of* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(\alpha)} - \boldsymbol{\beta})$ *given the* $\boldsymbol{e}_i$*'s with* $i \in \alpha$ *converges weakly to a multivariate normal distribution with zero mean and* $c\Sigma^{1/2}\mathbf{V}(\alpha)\Sigma^{1/2} \otimes \mathbf{Q}^{-1}$ *as the dispersion matrix where* c *and* $\mathbf{V}(\alpha)$ *are as in Theorem 4.2.2.*

Proof: Proof of this corollary follows from observing the fact that $n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$ converges to a positive definite matrix $\mathbf{Q}$ implies that $\max_{1\le i\le n}\boldsymbol{x}_i^T\{\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^T\}^{-1}\boldsymbol{x}_i$ converges to zero as n tends to infinity. □

Under the assumptions of Corollary 4.2.3, the expression of the asymptotic generalized variance becomes

$$[c^d\{\det(\Sigma)\}\det\{\mathbf{V}(\alpha)\}]^k[\det(\mathbf{Q})]^{-d}$$

The following Fact, which directly follows from Theorem 2.2.3 of Chapter 2 establishes a lower bound for $\det\{\mathbf{V}(\alpha)\}$ $(= v(\alpha)$, say), and this yields a lower bound for conditional asymptotic generalized variance of $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$.

Fact 4.2.4 *For the positive definite matrix* $\mathbf{D}(\alpha)$ *and the matrix* $\mathbf{J}(\alpha)$ *defined above, we have* $\det\{\mathbf{D}(\alpha)\} \ge [\det\{\mathbf{J}(\alpha)\}]^2$ *so that* $\det\{\mathbf{V}(\alpha)\} \ge 1$. *This lower bound is sharp in the sense that an exact equality in place of the inequality will hold if* $\mathbf{J}(\alpha)$ *happens to be an orthogonal matrix.*

We now propose to choose that subset α, which minimizes the asymptotic generalized variance of $\hat{\beta}_n^{(\alpha)}$. The above mentioned expression for generalized variance involves the scatter matrix Σ, which is in general unknown. We will need a consistent affine equivariant estimate $\hat{\Sigma}$ of Σ, and then we can transform y_i's to $\hat{\Sigma}^{-1/2}y_i$ for $1 \leq i \leq n$ and construct the transformation matrix $\mathbf{E}(\alpha)$ and the corresponding matrix $\hat{\mathbf{V}}(\alpha)$ as well as $\det\{\hat{\mathbf{V}}(\alpha)\}$ $(= \hat{v}(\alpha)$, say) based on those transformed observations. An optimal α is defined as $\hat{\alpha} = \arg\min_\alpha \hat{v}(\alpha)$. We now indicate the basic computational steps involved in the computation of the adaptive transformation retransformation estimate in multivariate median regression. From now on, we shall use the abbreviation TREMMER (Transformation Retransformation Estimate in Multivariate MEdian Regression) for that estimate.

4.2.2 TREMMER Algorithm

Step 1: Obtain a consistent and affine equivariant estimate $\hat{\Sigma}$ of the scale matrix Σ associated with the distribution of the random error e from the data $(y_1, x_1), \ldots, (y_n, x_n)$.

Step 2: Transform all the response vectors y_i's to $\hat{\Sigma}^{-1/2}y_i$'s for $1 \leq i \leq n$. Then fix a subset $\alpha \in S_n$ and compute $\hat{v}(\alpha)$ as given above which appears in the expression for asymptotic generalized variance of the estimate $\hat{\beta}_n^{(\alpha)}$.

Step 3: Minimize $\hat{v}(\alpha)$ with respect to $\alpha \in S_n$. Call that $\hat{\alpha}$. One can reduce the amount of computational time required for searching the optimal α by stopping whenever $\hat{v}(\alpha)$ is sufficiently close to 1 because we know from Fact 4.2.4 that the lower bound for $v(\alpha)$ is 1. This approximation makes the algorithm very fast.

Step 4: Form the matrix $\mathbf{W}(\hat{\alpha})$ with columns $x_{i_1}, \ldots, x_{i_k}$ where $i_1, \ldots, i_k$ are the first k elements of the subset $\hat{\alpha}$ and also form the matrix $\mathbf{Z}(\hat{\alpha})$ whose columns are $y_{i_1}, \ldots, y_{i_k}$. Then construct the transformation matrix $\mathbf{E}(\hat{\alpha})$ with columns $y_{j_1} - \mathbf{Z}(\hat{\alpha})\{\mathbf{W}(\hat{\alpha})\}^{-1}x_{j_1}, \ldots, y_{j_d} - \mathbf{Z}(\hat{\alpha})\{\mathbf{W}(\hat{\alpha})\}^{-1}x_{j_d}$ where $j_1, \ldots, j_d$ are the last d elements of $\hat{\alpha}$.

Step 5: Transform all response vectors y_i's to $z_i^{(\hat{\alpha})} = \{\mathbf{E}(\hat{\alpha})\}^{-1}y_i$ for $i \notin \hat{\alpha}$. Compute coordinate-wise LAD estimate $\hat{\Gamma}_n^{(\hat{\alpha})}$ of the matrix of parameters by regressing the $z_i^{(\hat{\alpha})}$'s on the x_i's for $i \notin \hat{\alpha}$. Then retransform that matrix to obtain the TREMMER estimate as $\hat{\beta}_n^{(\hat{\alpha})} = \hat{\Gamma}_n^{(\hat{\alpha})}\{\mathbf{E}(\hat{\alpha})\}^T$.

Before we discuss some applications of the TREMMER algorithm with real data sets, let us note that while transforming the response vectors by the square root of the variance covariance matrix computed from some preliminary error estimates is a popular approach (see Zellner 1962), the resulting coordinate system does not have any simple and natural geometric interpretation. Moreover, such a transformation does not lead to an affine equivariant modification of coordinatewise LAD estimates, and the limitation of such approach is primarily due to the fact that there does not exist an affine equivariant square

root of the usual estimates of the Σ matrix. Our 'data driven coordinate system' is a widely applicable tool for converting non-equivariant (or non-invariant) procedures into equivariant (or invariant) procedures, which is not limited to coordinatewise LAD estimates. Besides, for a properly selected subset α (as suggested in TREMMER algorithm) the matrix $[\mathrm{E}(\alpha)][\mathrm{E}(\alpha)]^T$ provides an estimate of the scale matrix Σ upto some scalar multiple.

In Step 1, we have to use a consistent and affine equivariant estimate of the scale matrix Σ. As the methodology is quite a general one, one may use any estimate with those properties and it is upon the user to select a proper estimate for his/her problem. Depending on the nature of the problem, one may use robust estimates of Σ (see, Davies 1987), but in general the construction of such robust estimates of Σ is computationally expensive, and if it is not absolutely necessary, one may use the variance covariance matrix of ordinary least squares residuals as an affine equivariant, consistent estimate of Σ.

Note that once the matrix $\mathrm{E}(\hat{\alpha})$ is formed, the computation of $\hat{\beta}_n^{(\hat{\alpha})}$ is straightforward as it requires to solve a linear programming problem for which a lot of efficient algorithms are available (Armstrong and Kung 1978, Barrodale and Roberts 1973, Wesolowsky 1981). As a result, the adaptive version of the TREMMER estimate continues to remain computationally advantageous. To compute the finite sample conditional variation of the TREMMER estimate given a fixed choice of the transformation, we have used resampling techniques like the bootstrap. To implement the bootstrap, one chooses the transformation matrix adaptively first, and then fixing that transformation matrix, one transforms all the y_i's to get the $z_i^{(\hat{\alpha})}$'s as before. Then one computes $\hat{\Gamma}_n^{(\hat{\alpha})}$ and retransforms it to get $\hat{\beta}_n^{(\hat{\alpha})}$. The sampling variation of $\hat{\beta}_n^{(\hat{\alpha})}$ is estimated by resampling from the pairs (y_i, x_i)'s for $1 \leq i \leq n$, $i \notin \hat{\alpha}$ and calculating the TREMMER estimate of β for each bootstrap replication keeping the optimal subset $\hat{\alpha}$ fixed. Then one computes the sample variance covariance matrix of those TREMMER estimates corresponding to different bootstrap samples. We next illustrate the procedure with Examples 4.1 and 4.2.

Example 4.1 (Continued): The following table gives the TREMMER estimates of regression coefficients and corresponding standard errors of the estimates are reported in the parentheses. Standard errors have been computed based on 10,000 bootstrap replications. In addition to the adaptive equivariant estimate, we have computed the nonequivariant least absolute deviation estimates of the regression parameters and estimated the generalized variances of both of them in order to make comparison. To compare two multidimensional estimates, Bickel (1964) defined the measure of efficiency as the p-th root of the ratio of corresponding generalized variances, where p is the dimension of the estimate. In the above example the dimension of the parameter is 4, and to compute the

Table 4.3: TREMMER estimates

Pressures	Constant	Age
Systolic	102.8509	0.8519
	(5.8851)	(0.2628)
Diastolic	73.1056	0.3587
	(3.6855)	(0.1425)

efficiency of TREMMER estimate we have taken the 4-th root of the ratio of the generalized variances of the TREMMER estimate and coordinate-wise LAD estimate. The efficiency estimated from 10,000 bootstrap replications turns out to be 1.145365. Figure 4.4 shows the TREMMER lines on the scatter plots of systolic and diastolic pressures against the age.

Example 4.2 (Continued): Total fertility rate (TFR) is a measure of fertility that denotes the average number of children born to a woman in her entire reproductive span assuming that she experiences the level of age-specific fertility rate obtained in a given year or period. Infant mortality rate (IMR) is defined as the number of deaths of children below age one year per 1000 live births. Detailed studies of the demographic transition in the developed and developing countries have revealed a strong link between declines in the mortality levels of a population (especially in the infant and child mortality) and fertility levels. One of the major determinants of demographic transition leading to decline in infant mortality and fertility is education of women. Female literacy rate (FLR) is defined as the percentage of literates among females aged 7 years and above. Our interest is to see the effect of FLR and time on TFR and IMR. The following table gives TREMMER estimates and corresponding standard errors of various regional effects and the effects of time and FLR in an analysis of covariance type linear model.

From Table 4.4, we see that both of FLR and time have strong negative effects on both TFR and IMR. So in India, infant mortality and fertility levels both seem to be declining with time and as female literacy increases. However, there is not much visible regional variation in the data, except for the fact that southern region tends to have the lowest TFR and IMR levels compared to others. In this example, we again observed that the TREMMER estimate is more efficient than the coordinate-wise LAD estimates the efficiency being 1.456471 as estimated from 10,000 bootstrap replications.

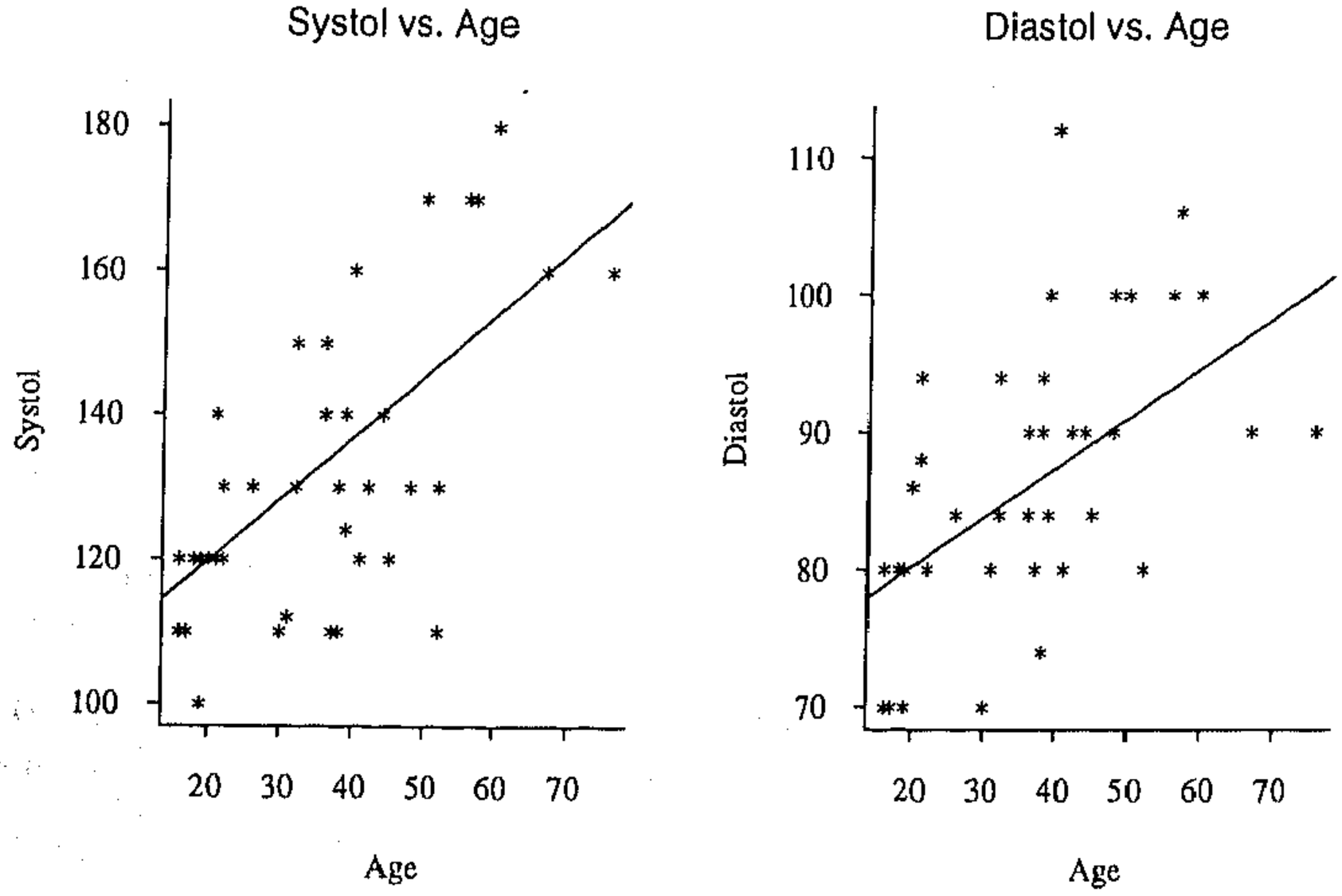


Figure 4.4: The plot shows TREMMER regression lines on the scatter plots of blood pressures with the age.

4.3 Asymptotic Optimality Properties of TREMMER

In this section, we will discuss asymptotic performance of the adaptive TREMMER estimate and establish some efficiency results. For that we impose some conditions on the x_i's.

Condition A : There exists a constant $M > 0$, a sequence of integers $\{k_n\}$ such that $k_n \to \infty$ as $n \to \infty$, and at least one partition of the set $\{1, 2, \ldots, n\}$ containing k_n subsets such that in each subset of that partition there exists at least one $\alpha \in S_n$ satisfying $\|\{\mathbf{W}(\alpha)\}^{-1} x_i\| \leq M$ for all $i \in \alpha$ and all n sufficiently large.

Condition B : The density h of a d-dimensional random vector e is spherically symmetric and satisfies

$$\int_{\mathbf{R}^{d\times k}} \{h(\sum_{i=1}^{k} a_i e_i)\}^d \prod_{i=1}^{k} h(e_i)\, de_i < \infty$$

where $e_1, \ldots, e_k$ are independent and identically distributed with common density h and

Table 4.4: TREMMER estimates for demographic data

	Regional Effects				Coeff. of	Coeff. of
	North	East	West	South	Time	FLR
TFR	6.9603	6.7627	6.7958	6.0459	-0.6361	-0.0338
	(0.3631)	(0.3589)	(0.3358)	(0.3640)	(0.1946)	(0.0139)
IMR	156.7407	164.0223	162.5226	144.9509	-10.5994	-1.2508
	(12.7284)	(19.5363)	(14.9983)	(14.7764)	(4.4422)	(0.4476)

a_i's are given constants.

Note that if the spherically symmetric density h is bounded, Condition B is trivially satisfied.

Remark : In the case of one-way analysis of variance problem, one can always construct a partition of the index set $\{1, 2, \ldots, n\}$ such that in each subset at least one replication of each treatment occurs. Note that in order to satisfy the condition $\max_{1\leq i\leq n} \boldsymbol{x}_i^T\{\sum_{j=1}^{n} \boldsymbol{x}_j\boldsymbol{x}_j^T\}^{-1}\boldsymbol{x}_i$ $\to 0$ as $n \to \infty$, the number of replications of each treatment goes to infinity. Thus one can easily have a sequence of partitions so that Condition A holds. We discuss in the following proposition another simple situation where Condition A holds.

Proposition 4.3.1 *Suppose that the $\boldsymbol{x}_i$'s are independent and identically distributed random variables satisfying $n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \xrightarrow{p} \mathbf{Q}$ as $n \to \infty$. Then the probability of the event that Condition A holds goes to one as n tends to infinity.*

Proof: As the $\boldsymbol{x}_i$'s are independent and identically distributed, there exists $M > 0$ such that for any $\alpha \in S_n$,

$$Pr[\max_{i\in\alpha} \|\{\mathbf{W}(\alpha)\}^{-1}\boldsymbol{x}_i\| < M] \equiv \delta > 0$$

for some $\delta > 0$. Consider any sequence of integers $\{k_n\}$ such that $k_n \longrightarrow \infty$ and $n/k_n \longrightarrow \infty$ as $n \longrightarrow \infty$. Then

$$Pr\{\text{ Condition A holds}\} \geq 1 - k_n(1-\delta)^{c_n/(k+d)}$$

where $c_n = n/k_n$. Thus the result follows immediately. □

Suppose that $\alpha^* \in S_n$ minimizes $\det\{\mathbf{V}(\alpha)\}$ $(= v(\alpha)$, say) which is defined in Theorem 4.2.2, when the scale matrix Σ is known.

Theorem 4.3.2 *Assume that the e_i's are independent and identically distributed with a common elliptically symmetric distribution $\{\det(\Sigma)\}^{-1/2} f(e^T\Sigma^{-1}e)$ such that the spherically symmetric density $h(e) = f(e^Te)$ on $\mathbb{R}^d$ satisfies Condition B, any univariate marginal g of h is differentiable and positive at 0, Σ is a $d \times d$ positive definite matrix, and the x_i's satisfy Condition A. Then $v(\alpha^*)$ converges to one in probability as n tends to infinity.*

Proof: First observe that in view of affine equivariance of $\hat{\beta}_n^{(\alpha)}$, it is enough to consider the case when $\Sigma = I_d$. Consider $A_{1n}, A_{2n}, \ldots, A_{k_n,n}$ disjoint subsets of $\{1, 2, \ldots, n\}$, such that Condition A holds. So for sufficiently large n, we will have at least one subset of indices $\alpha_i \in A_{in}$ such that $\|\{\mathbf{W}(\alpha_i)\}^{-1}x_{j_l}\|$ is bounded by M for $l = 1, \ldots, d$ and $\{j_1, j_2, \ldots, j_d\} \subset \alpha_i$. Note that for a subset of indices α, any column of the transformation matrix $\mathbf{E}(\alpha)$ can be written as $e_{j_1} - \sum_{l=1}^{k}(w_l^T x_{j_1})e_{i_l}$, where w_l^T is the l-th row of $\{\mathbf{W}(\alpha)\}^{-1}$. As the e_i's are i.i.d. with spherically symmetric density h, the joint p.d.f of $e_{i_1}, \ldots, e_{i_k}, e_{j_1}, \ldots, e_{j_d}$ can be written as $\prod_{i\in\alpha} h(e_i)$. Consider the following transformation of variables

$$u_1 = e_{j_1} - \sum_{l=1}^{k}(w_l^T x_{j_1})e_{i_l}, \ldots, u_d = e_{j_d} - \sum_{l=1}^{k}(w_l^T x_{j_d})e_{i_l}$$

$$u_{d+1} = e_{i_1}, \ldots, u_{d+k} = e_{i_k}$$

Then the joint density of $u_1, \ldots, u_{d+k}$ is given by

$$\prod_{i=1}^{d} h\{u_i + \sum_{l=1}^{k}(w_l^T x_{j_i})u_{d+l}\} \prod_{i=1}^{k} h(u_{d+i})$$

Therefore, the joint density of $u_1, \ldots, u_d$ at the origin in $\mathbb{R}^{d\times d}$ is

$$\int_{\mathbb{R}^{d\times k}} \prod_{i=1}^{d} h\{\sum_{l=1}^{k}(w_l^T x_{j_i})u_{d+l}\} \prod_{i=1}^{k} h(u_{d+i})\, du_{d+1} \ldots du_{d+k},$$

which exists and is positive by Condition B. Now, in view of Condition A and the continuity of h at $0 \in \mathbb{R}^d$, the joint density of $u_1, \ldots, u_d$ must remain bounded away from zero in a neighborhood of $0 \in \mathbb{R}^{d\times d}$. Therefore the probability that the columns of $\mathbf{E}(\alpha)$ will be near orthogonal (and hence $v(\alpha) = \det\{\mathbf{V}(\alpha)\}$ will be very close to 1) is bounded away from zero. So we have for any $\epsilon > 0$

$$\inf_{x_i; i\in\alpha} Pr[v(\alpha) < 1 + \epsilon] = p_\epsilon > 0$$

Then

$$\begin{aligned} Pr\{v(\alpha^*) \geq 1 + \epsilon\} &= Pr\{\forall \alpha \in S_n, \quad v(\alpha) \geq 1 + \epsilon\} \\ &\leq Pr\{v(\alpha_1) \geq 1 + \epsilon, \ldots, v(\alpha_{k_n}) \geq 1 + \epsilon\} \\ &\leq (1 - p_\epsilon)^{k_n} \longrightarrow 0 \ \ as \ \ n \to \infty. \end{aligned}$$

□

The above Theorem implies that if the scale matrix Σ happens to be known and the adaptive selection of $\alpha^* \in S_n$ is done using that known Σ, the conditional generalized variance of the resulting adaptive TREMMER estimate tends to the lower bound established in Fact 4.2.4. However, in practice Σ is unknown, and we will estimate it by a consistent and affine equivariant estimate $\hat{\Sigma}$ when we minimize $\hat{v}(\alpha)$ to obtain $\hat{\alpha}$. The next Theorem tells that the difference between $v(\hat{\alpha})$ and $v(\alpha^*)$ is asymptotically negligible.

Theorem 4.3.3 *Under the assumptions of the previous theorem, $v(\hat{\alpha}) - v(\alpha^*)$ converges in probability to zero as n tends to infinity.*

Proof: As $\hat{\Sigma}$ is a consistent estimate of Σ, by the simple arguments used in the proof of Lemma 2.2.7, 2.2.8 in Chapter 2, it can be shown that $\sup_{\alpha \in S_n} |\hat{\mathbf{J}}(\alpha) - \mathbf{J}(\alpha)|$ converges *in probability* to zero as n tends to infinity, which in turn implies that

$$\sup_{\alpha \in S_n} |\hat{\mathbf{D}}(\alpha) - \mathbf{D}(\alpha)| \xrightarrow{p} 0, \tag{4.3}$$

$$\sup_{\alpha \in S_n} |[\det\{\hat{\mathbf{J}}(\alpha)\}]^2 - [\det\{\mathbf{J}(\alpha)\}]^2| \xrightarrow{p} 0 \tag{4.4}$$

and

$$\sup_{\alpha \in S_n} |\det\{\hat{\mathbf{D}}(\alpha)\} - \det\{\mathbf{D}(\alpha)\}| \xrightarrow{p} 0 \quad as \quad n \longrightarrow \infty. \tag{4.5}$$

For $M' > 0$, define $K^n_{M'} = \{\alpha : \alpha \in S_n \text{ and } v(\alpha) \leq M'\}$. Then by (4.3), (4.4) and (4.5) it is easy to see that $\sup_{\alpha \in K^n_{M'}} |\hat{v}(\alpha) - v(\alpha)|$ converges *in probability* to zero as n tends to infinity.

From Theorem 4.3.2, we have that the α^*, which minimizes $v(\alpha)$, is in the set $K^n_{M'}$, and hence in view of the fact stated above $\hat{\alpha}$ will be in $K^n_{M'}$ with probability tending to one as n tends to infinity if $M' > 0$ is chosen to be suitably large.

Next, since $\hat{\alpha}$ minimizes $\hat{v}(\alpha)$, and α^* minimizes $v(\alpha)$, it follows by some straightforward analysis that $|\hat{v}(\hat{\alpha}) - v(\hat{\alpha})| < \epsilon$ and $|\hat{v}(\alpha^*) - v(\alpha^*)| < \epsilon$ will imply that $|\hat{v}(\hat{\alpha}) - v(\alpha^*)| < \epsilon$. Hence, it follows that $\hat{v}(\hat{\alpha}) - v(\alpha^*)$ converges *in probability* to zero, which completes the proof with previous observations. □

The above two theorems suggest that there is an optimal choice of the subset $\alpha \in S_n$ for which $v(\alpha)$ attains its lower bound as n goes to infinity when the scale matrix Σ is known. If Σ is unknown, with a consistent and equivariant estimate of Σ, we can choose a subset $\hat{\alpha} \in S_n$ such that $v(\hat{\alpha})$ also attains its lower bound asymptotically. Thus for n sufficiently large, we will be able to get hold of an $\hat{\alpha}$ such that $v(\hat{\alpha}) < 1 + \epsilon$, for any $\epsilon > 0$. Any $\alpha \in S_n$, for which $v(\alpha) < 1 + \epsilon$, will produce an estimate with conditional asymptotic generalized variance close to $[(c/n)^d \det(\Sigma)]^k [\det(\mathbf{Q})]^{-d}$, where $\mathbf{Q}$ is the positive definite

limit of $n^{-1}\sum_{i=1}^{n} x_i x_i^T$ as n tends to infinity. From the asymptotic result obtained by Rao (1988), it can be seen that the asymptotic generalized variance of the coordinate-wise LAD estimates of parameter matrix is $[(c/n)^d \det(\Gamma)]^k [\det(\mathbf{Q})]^{-d}$, where the (i,j)-th element of Γ is $(\sigma_{ii}\sigma_{jj})^{1/2}(2/\pi)\sin^{-1}\rho_{ij}$, $\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2}$. Here σ_{ij} is the (i,j)-th element of Σ and c is as defined earlier. Following the line of arguments used in the proof of Fact 2.2.3 in Chapter 2, it is easy to see that $\det(\Gamma) \geq \det(\Sigma)$, and equality holds only if Σ is a diagonal matrix. If the asymptotic efficiency of two competing estimates of the kd-dimensional parameter matrix is defined as the $(k \times d)$-th root of the ratio of their asymptotic generalized variances, the efficiency of TREMMER estimate compared to nonequivariant coordinate-wise LAD estimate is always greater than or equal to one. Further, Theorems 4.3.2 and 4.3.3 imply that it is possible to get hold of an appropriate transformation matrix $\mathbf{E}(\alpha)$ for large n such that the estimate $\hat{\beta}_n^{(\alpha)}$ will be more (or less) efficient than the ordinary least squares estimate depending on whether the tail of the univariate marginal g of the spherically symmetric density $f(e^T e)$ is 'heavy' (or 'light'). Observe that we are using a linear transformation which retains the linear structure of the model, and the efficiency gain is solely due to non-equivariance of the coordinatewise LAD estimates in multiresponse linear models under nonsingular linear transformations.

We close this section by presenting some simulation results to demonstrate the performance of the adaptive TREMMER estimate in small samples. In the model $y_i = \beta^T x_i + e_i$, we have generated the e_i's from bivariate normal (i.e. $f(e^T e) = (2\pi)^{-1}\exp(-(e^T e)/2)$) and Laplace (i.e. $f(e^T e) = (2\pi)^{-1}\exp(-\sqrt{e^T e})$) distributions with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We have taken $\beta = 0$, $k = 2$, where the first element of x_i is one and the second element is generated from standard univariate normal distribution. Using these e_i's, x_i's and β, we have generated the observations (y_i, x_i)'s for $i = 1, \ldots, n$. We have used a set of five different values of ρ and two sample sizes, namely 20 and 30. Our adaptive TREMMER estimate was compared with the coordinate-wise LAD estimate, and for the purpose of efficiency computation, the estimates of their generalized variances were obtained based on 10,000 Monte Carlo replications. The efficiency is taken to be the fourth root of the ratio of the generalized variances of the two competing estimates of β.

From Tables 4.5 and 4.6, we see that TREMMER estimates are more efficient than coordinate-wise LAD estimates in the presence of substantial correlations even with small sample sizes. As correlation among the real valued coordinates of the response vector increases, the efficiency of TREMMER over coordinate-wise LAD increases. It will be appropriate to note here that unlike what has been done in Examples 1 and 2 where

we estimated conditional sampling variation using the bootstrap, in these simulations we have compared unconditional sampling variation of TREMMER estimates with that of the coordinate-wise LAD.

Table 4.5: Efficiency figures for bivariate normal

Sample Size	ρ				
	0.75	0.80	0.85	0.90	0.95
20	1.0539	1.2695	1.4931	1.3253	1.8995
30	1.2391	1.2590	1.2251	1.5327	1.9898

Table 4.6: Efficiency figures for bivariate Laplace

Sample Size	ρ				
	0.75	0.80	0.85	0.90	0.95
20	1.0431	1.1825	1.3816	1.4899	1.6065
30	1.0181	1.2396	1.4935	1.6243	1.6740

We conclude by noting that when the underlying distribution of the e_i's are not elliptically symmetric, the conditional asymptotic normality of $\hat{\beta}_n^{(\alpha)}$ still holds but with a more complicated dispersion matrix. To choose the best subset α in that case, one can estimate the asymptotic generalized variance of $\hat{\beta}_n^{(\alpha)}$ for a given α by resampling or some other technique and then try to minimize that over different possible choices of α. But that will be much more computationally intensive, and we do not intend to consider it here.

4.4 Transformation Retransformation Procedure and Multivariate Rank Regression

Let us now focus our attention on the data points (x_i, y_i)'s, which are assumed to satisfy the linear model (4.1). Suppose that $n > d + k$, and α is a subset of size $d + k$ of the set of indices $\{1, 2, \ldots, n\}$. Following the notation used in the previous Sections, we will write $\alpha = \{i_1, \ldots, i_k, j_1, \ldots, j_d\}$ and denote by $\mathbf{W}(\alpha)$ the $k \times k$ matrix whose columns are the vectors $x_{i_1}, \ldots, x_{i_k}$ and by $\mathbf{Z}(\alpha)$ the $d \times k$ matrix whose columns are the vectors $y_{i_1}, \ldots, y_{i_k}$. We will again assume that $\mathbf{W}(\alpha)$ is invertible and form the $d \times d$ matrix $\mathbf{E}(\alpha)$ that consists of the columns $y_{j_1} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}x_{j_1}, \ldots, y_{j_d} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}x_{j_d}$. The matrix $\mathbf{E}(\alpha)$ too is assumed to be non-singular, and as before we define the transformed response vectors as $z_l^{(\alpha)} = \{\mathbf{E}(\alpha)\}^{-1}y_l$ for $1 \leq l \leq n$ and $l \notin \alpha$. Suppose now that we

perform rank regression on each coordinate of $z_l^{(\alpha)}$ separately with x_l as the regressor as has been done in Davis and McKean (1993), and the resulting estimate of the matrix of coefficient parameters is denoted by $\hat{\Lambda}_n^{(\alpha)}$. In other words, $\hat{\Lambda}_n^{(\alpha)}$ is obtained by minimizing (w.r.t. Λ) a dispersion function $\mathcal{D}(\Lambda)$ (say), which is a simple multivariate extension of Jaeckel's dispersion function [see Jaeckel (1972)] based on residuals and their ranks computed from a linear model. In this case $\mathcal{D}(\Lambda)$ is a function of the real valued coordinates of the multivariate residuals $z_l^{(\alpha)} - \Lambda^T x_l$ with $1 \leq l \leq n$, $l \notin \alpha$ and their ranks [see Davis and McKean (1993)]. Finally, the transformation retransformation estimate of β is obtained by retransforming $\hat{\Lambda}_n^{(\alpha)}$ by the matrix $\mathbf{E}(\alpha)$ as follows

$$\hat{\beta}_n^{(\alpha)} = \hat{\Lambda}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^T. \tag{4.6}$$

In view of the definition of $\hat{\beta}_n^{(\alpha)}$, we now have the following result, which asserts that it is an affine equivariant estimate of β. As a matter of fact, this result is the analog of Theorem 4.2.1 in the context of rank regression.

Result 4.4.1 *Suppose that* $\mathbf{A}$ *is a fixed* $d \times d$ *nonsingular matrix. Then the transformation retransformation estimate computed from* $(\mathbf{A}y_1, x_1), (\mathbf{A}y_2, x_2), \ldots, (\mathbf{A}y_n, x_n)$ *in the same way as above (i.e. using the same index set* α*) will be* $\hat{\beta}_n^{(\alpha)}\mathbf{A}^T$*. Further, if the response vector* y_i *is transformed to* $y_i - \mathbf{G}^T x_i$ *for each* $i = 1, 2, \ldots, n$*, where* $\mathbf{G}$ *is a fixed* $d \times k$ *matrix, the transformation retransformation estimate gets transformed to* $\hat{\beta}_n^{(\alpha)} - \mathbf{G}$.

4.4.1 Selection of the Optimal Data Driven Transformation

Since the estimate $\hat{\beta}_n^{(\alpha)}$ depends on the choice of the transformation matrix $\mathbf{E}(\alpha)$, a question that naturally arises at this point is how to choose the subset of indices α. Depending on the nature of the problem, we have earlier determined the form of the optimal transformation $\mathbf{E}(\alpha)$ and suggested appropriate data driven selection procedure for the optimal subset of indices α. All these procedures for choosing the optimal transformation matrix, however, are based on the common idea of minimizing the generalized variance (i.e. the determinant of the dispersion matrix) of the multivariate location or regression estimate. The motivation for looking at the generalized variance comes from the fact that it is proportional to the volume of the concentration ellipsoid associated with the sampling distribution of the estimate which is usually normal for large samples. We will now state a result that asserts that under suitable regularity conditions $\hat{\beta}_n^{(\alpha)}$ is a $n^{1/2}$-consistent and asymptotically normal estimate of the parameter matrix β in the linear model (4.1).

Result 4.4.2 *Fix an* α*. Suppose that the distribution of the* (x_i, y_i)*'s and the nature of the dispersion function* $\mathcal{D}(\Lambda)$ *are such that* $n^{1/2}$*-consistency and asymptotic normality of*

the coordinatewise rank regression estimates holds. For example the regularity conditions used in Davis and McKean (1993), who considered coordinatewise rank regression will be sufficient for this purpose. Then conditioned on α *and the* (x_i, y_i)*'s with* $i \in \alpha$*, the asymptotic distribution of* $n^{1/2}(\hat{\beta}_n^{(\alpha)} - \beta)$ *is multivariate normal with zero mean and a variance covariance matrix that depends on the transformation matrix* $\mathbf{E}(\alpha)$.

Proof: Let us fix an α and argue conditionally given the (x_i, y_i)'s with $i \in \alpha$. Note that since $\hat{\Lambda}_n^{(\alpha)}$ is obtained by performing coordinatewise rank regression of the transformed response vectors $z_l^{(\alpha)}$'s on the covariates x_l's, it will be a $n^{1/2}$-consistent and asymptotically normal estimate of $\{\mathbf{E}(\alpha)\}^{-1}\beta^T$ under appropriate regularity conditions as assumed in the statement of the result. The proof is now complete if we recall that $\hat{\beta}_n^{(\alpha)} = \hat{\Lambda}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^T$ and use the fact that linear transformation preserves multivariate normality as well as $n^{1/2}$-consistency. □

However, the conditional asymptotic dispersion matrix of $\hat{\beta}_n^{(\alpha)}$ depends on $\mathbf{E}(\alpha)$ in a rather complex way, and it is hardly useful in providing any insight regarding the optimal choice of α in a general situation. Alternatively, one can try to use resampling techniques (e.g. the bootstrap) to estimate the sampling variation in $\hat{\beta}_n^{(\alpha)}$, and then select an optimal $\mathbf{E}(\alpha)$ based on this estimate. But, it does not seem to be a feasible approach in practice in view of the enormous amount of computation that any form of resampling estimation of the dispersion of $\hat{\beta}_n^{(\alpha)}$ will require for different choices of α.

Suppose now that e has an elliptically symmetric distribution with a density of the form $\{\det(\Sigma)\}^{-1/2} f(e^T\Sigma^{-1}e)$, where Σ is a $d \times d$ positive definite matrix, and f is a probability density function on the real line. Let us write $\{\Sigma^{-1/2}\mathbf{E}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$, where $\mathbf{R}(\alpha)$ is a diagonal matrix with positive diagonal entries, and $\mathbf{J}(\alpha)$ is a matrix whose rows are of unit length, and define $\mathbf{D}(\alpha)$ to be the symmetric $d \times d$ matrix whose (i, j)-th element is $\sin^{-1}\gamma_{ij}$, γ_{ij} being the Euclidean inner product of the i-th and the j-th row of $\mathbf{J}(\alpha)$. Then by Theorem 4.2.2 under suitable conditions the asymptotic generalized variance of the transformation retransformation median regression (i.e. TREMMER) estimate of β in the linear model (4.1) is minimized by choosing α to minimize the determinant of the matrix

$$\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^T\}^{-1}. \tag{4.7}$$

Note that such a selection of α does not require any knowledge of the form of the density f, and there is a nice and intuitively appealing geometric interpretation for such an approach. The determinant of $\mathbf{V}(\alpha)$ is minimized when the columns of $\Sigma^{-1/2}\mathbf{E}(\alpha)$ are orthogonal to one another. Hence, an alternative way of selecting $\mathbf{E}(\alpha)$ will be to minimize the ratio of the trace and the d-th root of the determinant of the matrix $\{\mathbf{E}(\alpha)\}^T\Sigma^{-1}\mathbf{E}(\alpha)$, which

is equivalent to minimizing the ratio of the arithmetic mean and the geometric mean of the eigenvalues of the positive definite matrix [see Chapters 2 and 3]. In the absence of any other better and practically feasible procedure, we intend to use this criterion for choosing the transformation matrix for our multivariate rank regression. In other words, our recommendation amounts to transforming the response vectors using a new data driven coordinate system determined by the transformation matrix $\mathbf{E}(\alpha)$ such that the coordinate system is as orthogonal as possible in the d-dimensional vector space, where the inner product and orthogonality are defined based on the positive definite dispersion matrix Σ of the residual distribution associated with the linear model (4.1). Of course we need an appropriate estimate of Σ in order to implement such a strategy, and we can get that from the residuals computed at an initial stage after fitting the linear model to the data by any simple and suitable method. Note that it is important that such an estimate of Σ be equivariant under linear transformation of the response vectors.

4.4.2 Multivariate Rank Regression Using Wilcoxon's Score

Let us now consider the dispersion functions associated with well known Wilcoxon's rank scores. Such dispersion functions can be expressed in the form

$$\mathcal{D}(\Lambda) = \sum_{1\le r<s\le n;\, r,s\notin\alpha} \sum \left|(z_r^{(\alpha)} + z_s^{(\alpha)}) - \Lambda^T(x_r + x_s)\right| \tag{4.8}$$

or

$$\mathcal{D}(\Lambda) = \sum_{1\le r<s\le n;\, r,s\notin\alpha} \sum \left|(z_r^{(\alpha)} - z_s^{(\alpha)}) - \Lambda^T(x_r - x_s)\right| , \tag{4.9}$$

where for a d-dimensional vector $x = (x_1, x_2, \ldots, x_d)$, $|x|$ = the l_1-norm of $x = |x_1| + |x_2| + \ldots + |x_d|$. Note that the dispersion in (4.8) originates from Wilcoxon's signed rank score used in single sample location problems while that in (4.9) is related to the two sample Wilcoxon's rank test. The second dispersion can be viewed as a form of Gini's mean difference of multivariate residuals, and it is meaningful to use this dispersion function when there is no intercept term present in the linear model (4.1). On the other hand the dispersion function in (4.8) is useful in multivariate linear models with intercept terms. Readers are referred to Aubuchon and Hettmansperger (1989) and Chaudhuri (1992b) for a detailed discussion of these dispersion functions and their use in rank regression in linear models with univariate response.

The estimates of the coefficient matrix obtained through minimization of dispersion functions in (4.8) and (4.9) can be viewed as natural extensions of the well-known Hodges-Lehmann estimates from one and two sample location problems into multivariate linear models. Observe at this point that minimization of any of these two dispersions leads

to a coordinatewise least absolute deviations problem, and hence the computation of the transformation retransformation estimate $\hat{\beta}_n^{(\alpha)}$ can be easily handled by some straight forward modification of the TREMMER algorithm developed in Section 4.2.2. One only needs to replace the original data by their pairwise averages or differences (depending on whether (4.8) or (4.9) is used) before invoking TREMMER. We now state a result that establishes asymptotic optimality of our procedure for choosing the transformation matrix $\mathbf{E}(\alpha)$ as described in Section 4.4.1 when rank regression is performed using Wilcoxon's score in a multivariate linear model with the residual having multivariate normal distribution.

Result 4.4.3 *Suppose that the residuals $e_i = y_i - \beta^T x_i$ for $1 \leq i \leq n$ are i.i.d and have a common d-variate normal distribution with zero mean and Σ as their common dispersion matrix that does not depend on the regressor (i.e. we have perfect homoscedasticity), and the i.i.d random regressors x_i's have a distribution with an associated $k \times k$ expected information matrix $E(x_i x_i^T) = \mathbf{Q}$ that is positive definite ensuring asymptotic normality of the coordinatewise rank regression estimates obtained using the dispersion function (4.8) or (4.9) [cf. the asymptotic results in Chaudhuri (1992b)]. Then our procedure for choosing the set of indices α and the associated transformation matrix $\mathbf{E}(\alpha)$ described in Section 4.4.1 yields a transformation retransformation estimate $\hat{\beta}_n^{(\alpha)}$ such that the asymptotic generalized variance of $n^{1/2}(\hat{\beta}_n^{(\alpha)} - \beta)$ tends to its minimum possible value as n tends to infinity.*

Proof : Once again let us fix α and argue conditionally give the (x_i, y_i)'s with $i \in \alpha$. Note that when the dispersion function (4.9) is used, there are no intercept terms in the multivariate linear model, and without loss of generality we can assume in this case that the x_i's have zero mean. Under the conditions assumed in the statement of the result, it is easy to establish a Bahadur type asymptotic linear representation of $\hat{\beta}_n^{(\alpha)}$ using the asymptotic results in Chaudhuri (1992b), and this implies that as n tends to infinity, the limiting distribution of $n^{1/2}(\hat{\beta}_n^{(\alpha)} - \beta)$ is multivariate normal with zero mean and a variance covariance matrix that has the form

$$\Sigma^{1/2}\{\mathbf{J}(\alpha)\}^{-1}\mathbf{H}(\alpha)\{[\mathbf{J}(\alpha)]^T\}^{-1}\Sigma^{1/2} \otimes \mathbf{Q}^{-1}, \tag{4.10}$$

where $\otimes$ denotes the usual Kornecker product of matrices. Here $\mathbf{J}(\alpha)$ is the matrix whose rows are obtained by normalizing the rows of the matrix $\{\Sigma^{-1/2}\mathbf{E}(\alpha)\}^{-1}$ as described in Section 4.4.1, and $\mathbf{H}(\alpha)$ is the $d \times d$ symmetric matrix with (i,j)-th element equal to $2\sin^{-1}(\gamma_{ij}/2)$, γ_{ij} being the Euclidean inner product between the i-th and the j-th row of $\mathbf{J}(\alpha)$. It is the multivariate normality of the residual distribution in the linear model that enables us to simplify the the form of the asymptotic dispersion matrix in this special

case. It is clear from (4.10) that the asymptotic generalized variance of the transformation retransformation rank regression estimate will be minimized if we choose α to minimize $\det\{\mathbf{H}(\alpha)\}/[\det\{\mathbf{J}(\alpha)\}]^2$, and this is accomplished when the rows of $\mathbf{J}(\alpha)$ or equivalently the columns of $\Sigma^{-1/2}\mathbf{E}(\alpha)$ are orthogonal to one another. □

4.5 Numerical Results : Simulation and Data Analysis

In an attempt to investigate the performance of transformation retransformation rank reregression methodology in finite sample situations, we ran a simulation study and analyzed the real data sets in Examples 4.1 and 4.2 . We compared our approach with more traditional procedures some of which are not affine equivariant, and as we will gradually see the results turned out to be quite encouraging and favorable for our affine equivariant rank regression.

A Simulation Study : We considered a problem with sample size $n = 30$, where the data was generated from a multivariate linear model like (4.1) with $d = k = 2$, and the first coordinate of $\boldsymbol{x}$ was taken to be the constant 1.0 while the second coordinate was generated from a standard normal distribution. We chose $\boldsymbol{\beta}$ as the 2×2 zero matrix, and for the random residual, we used three different elliptically symmetric distributions i.e. distributions having densities of the form $\{\det(\Sigma)\}^{-1/2} f(\boldsymbol{e}^T\Sigma^{-1}\boldsymbol{e})$. These distributions are bivariate normal, bivariate Laplace [i.e. when $f(\boldsymbol{e}^T\boldsymbol{e}) = (2\pi)^{-1}\exp(\sqrt{\boldsymbol{e}^T\boldsymbol{e}})$] and bivariate t with 3 degrees of freedom. We used the dispersion function (4.8) for computing the transformation retransformation estimate $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ after choosing α using the selection procedure described in Section 4.4.1. Let $\mathcal{E}_{ols}$ and $\mathcal{E}_{lad}$ denote the efficiencies of our estimates compared with the ordinary least squares and coordinatewise least absolute deviations estimates respectively. These efficiencies were computed using the fourth root of the ratio of the generalized variances of competing estimates [see e.g. Bickel (1964)], and the generalized variances were estimated using 3000 Monte Carlo replications in each case. Since both of ordinary least squares estimate and our estimate of $\boldsymbol{\beta}$ are affine equivariant, $\mathcal{E}_{ols}$ does not depend on Σ. We observed that for bivariate normal $\mathcal{E}_{ols} = 82\%$, and for bivariate Laplace $\mathcal{E}_{ols} = 101\%$. However, for the t distribution with 3 degrees of freedom, which is a distribution with a fairly heavy tail, we observed that $\mathcal{E}_{ols} = 150\%$. Since the coordinatewise least absolute deviations regression does not lead to an affine equivariant estimate of $\boldsymbol{\beta}$, $\mathcal{E}_{lad}$ depends on Σ. For our simulation study, we have used different choices of Σ, and each choice had both diagonal entries equal to 1.0 and both off-diagonal entries equal to ρ. Five different values of ρ were used, and they are 0.75, 0.80, 0.85, 0.90 and 0.95. The results are summarized in Table 4.7.

Table 4.7: Values of $\mathcal{E}_{lad}$ for different choices of the residual distribution and ρ.

Residual Distribution	Values of ρ				
	0.75	0.80	0.85	0.90	0.95
Bivariate Normal	1.5823	1.7436	1.8547	2.0782	2.4439
Bivariate Laplace	1.2581	1.2874	1.4437	1.5925	1.9356
Bivariate t with 3 d.f.	1.2135	1.2843	1.4138	1.5829	1.8910

Example 4.1 (Continued): We applied our affine equivariant rank regression procedure based on the dispersion function (4.8) to the blood pressures data and obtained the following estimated linear equations : *systolic pressure* = 100.64 + 0.8(*age*), and *diastolic pressure* = 74.04 + 0.32(*age*). Following the procedures used earlier, we estimated the sampling variations using 2000 bootstrap samples for each of the competing procedures and observed 66.9% gain in statistical efficiency when our affine equivariant rank regression was compared with coordinatewise least absolute deviations regression. The coefficients of age in both the equations here are slightly larger than those obtained using TREMMER, and their standard errors (0.20 and 0.11 for systolic and diastolic pressures respectively) estimated through bootstrap turned out to be smaller than those for the TREMMER estimates.

Example 4.2 (Continued): Since here one is interested in the differences between regional effects, the dispersion function in (4.9) is quite appropriate. When we compared our affine equivariant procedure with non-equivariant coordinatewise rank regression based on Wilcoxon's score using botstrap estimates of sampling variations, we observed about 8% gain in statistical efficiency. As in the preceding example, here too we used 2000 bootstrap samples for each competing procedure. In the case of our affine equivariant procedure, time with estimated coefficients -0.4929 and -9.5899 having standard errors 0.1964 and 4.5880 respectively appeared to be a statistically significant covariate indicating decline in both of TFR and IMR over time. FLR too turned out to be a statistically significant covariate with estimated coefficients -0.03775 and -1.3006 having standard errors 0.01223 and 0.2983 respectively indicating a strong influence of female education on decreasing TFR and IMR. However, as in case of TREMMER based analysis we did not observe any statistically significant regional difference in fertility and mortality rates.

Chapter 5

Multivariate Quantiles

5.1 Introduction

The problem of finding out suitable analogs of quantiles for multivariate data has a long history in statistics. Univariate quantiles are quite popular for their usefulness in constructing useful descriptive statistics like the median, the inter-quartile range and various measures of skewness and kurtosis. They are also used in constructing robust L-estimates of location. As there is no inherent ordering in multidimension, extending the notion of quantiles poses a big problem. In a classic paper, Barnett (1976) reviewed different possible techniques for ordering multivariate data (see also Plackett, 1976 and Reiss, 1989). Brown and Hettmansperger (1987, 1989) have proposed a notion of bivariate quantiles based on Oja's simplicial median (see Oja, 1983). Eddy (1983, 1985) proposed an interesting approach to define multivariate quantiles using certain nested sequence of convex sets. Very recently, Chaudhuri (1996) and Koltchinskii (1997) proposed the notion of geometric or spatial quantile, which generalizes the notion of spatial median that has been studied by earlier authors (see e.g. Brown 1983, Chaudhuri 1992a). Chaudhuri (1996) indexed multivariate geometric quantiles, based on Euclidean distances, using the elements of d-dimensional open unit ball. The corresponding quantiles not only give the idea of 'extreme' or 'central' observations but also about their orientations in the data cloud. He also presented a Bahadur type representation for the geometric quantiles and indicated various ways of extending these quantiles to L-estimates, regression quantiles etc. Recently, Mardon (1998) proposed some analogs of bivariate Q-Q plots based on geometric quantiles. These bivariate Q-Q plots can be used in comparing a sample to a given population distribution and they may reveal differences in location, scale and skewness, as well as outliers.

Whether the notion of multivariate quantiles would be based on some univariate con-

cept of ordering or on some vector valued concept of ranks is a debatable issue. In many ways it seems to be a good idea to make use of the orientation information in any version of multivariate quantile. That is the only way in which one can talk about the 'high points' and the 'low points' in a multivariate data cloud. In a multivariate situation an observation may have 'high' values in some direction but 'low' values in some other direction. To capture these intrinsic geometric features of the multivariate data cloud, it seems reasonable to index the multivariate quantiles by some multivariate quantities, which will give us a way of measuring the closeness (or deviation) of a specific data point to (or from) the center of the data cloud as well as its spatial orientation with respect to the data cloud. Brown and Hettmansperger (1987, 1989) introduced a notion of bivariate quantile which is based on their definition of multivariate ranks derived from Oja's criterion function (cf. Oja, 1983). The problem with their approach is that the criterion function used by them is not 'self-normalized' in the sense that it is the gradient vector of Oja loss function based on simplicial volumes and is not bounded. For certain losses and distances, the gradient leads to 'self-normalized' orientation. For instance, the gradient vector of the function $f(x_1, x_2, \ldots, x_d) = |x_1| + |x_2| + \cdots + |x_d|$ (i.e. the l_1-norm) is the coordinatewise sign vectors for which each coordinate is bounded by 1. If $f(x_1, x_2, \ldots, x_d) = (x_1^2 + x_2^2 + \cdots + x_d^2)^{1/2}$ (i.e. the l_2-norm), the gradient is a unit direction vector [see Möttönen and Oja (1995) and Möttönen, Oja and Tienari (1997) for the notions of the ranks of the data points constructed using such a gradient]. The advantage of using 'self-normalized' orientation is that it becomes easy to interpret what is 'high' and what is 'low' in a multidimensional setting.

The problem with geometric quantiles (Chaudhuri 1996, Koltchinskii 1997) is that they are not equivariant under arbitrary affine transformations though they are equivariant under rotations of the data cloud. Due to lack of affine equivariance, these geometric quantiles do not lead to any sensible estimate when the different coordinate variables of the data-vectors are measured in different units or they have different degrees of statistical variations. In this Chapter we have used the transformation retransformation approach to construct affine equivariant estimates of multivariate quantiles. In Section 5.2, we introduce the notion of l_p-quantiles and a proper indexing for them. Then with the help of transformation retransformation methodology, we extend l_p-quantiles to a family of affine equivariant multivariate quantiles and explore their basic properties with regard to uniqueness, existence and computation. In Section 5.3, we discuss asymptotic behavior of multivariate quantiles. We establish a Bahadur-type linear representation and use it to derive asymptotic distributions of sample quantiles. In the same section, we indicate a procedure to select a suitable 'data-driven coordinate system' and discuss a few interesting results related to that. In Section 5.4, we present some applications of our proposed quantiles. In

particular, we discuss construction of quantile based contour plots for distributions and indicate a procedure for multivariate generalization of Q-Q plots and demonstrate with some simulation results and real data sets about how they can be used in comparing a multivariate sample to a given distribution. We also construct L-estimates and trimmed mean estimates for multivariate location based on these multivariate quantiles.

5.2 l_p-Quantiles and Transformation Retransformation

It is easy to see that given any β such that $0 < \beta < 1$ and $u = 2\beta - 1$, the sum $\sum_{i=1}^{n}\{|X_i - Q| + u(X_i - Q)\}$ is minimized when Q is the sample β-th quantile based on the real-valued observations X_i's (see e.g. Ferguson 1967). In this article, we generalize this concept to d-dimensional l_p spaces for $1 \leq p < \infty$. Define the open unit ball $B_p^{(d)}$ in l_p space as $\{\boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_p < 1\}$ where $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ and $\|\boldsymbol{u}\|_p = (|u_1|^p + \cdots + |u_d|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max(|u_1|, \ldots, |u_d|)$. For $1 \leq p < \infty$, and for any $\boldsymbol{u} \in B_q^{(d)}$, $\boldsymbol{t} \in \mathbb{R}^d$, where $1/p + 1/q = 1$ with the convention that $q = \infty$ when $p = 1$, let us define

$$\Phi_p(\boldsymbol{u}, \boldsymbol{t}) = \|\boldsymbol{t}\|_p + \boldsymbol{u}^T\boldsymbol{t}. \tag{5.1}$$

Then the l_p-quantile $\hat{Q}_n^{(p)}(\boldsymbol{u})$ corresponding to $\boldsymbol{u}$ is defined as

$$\hat{Q}_n^{(p)}(\boldsymbol{u}) = \arg\min_{Q \in \mathbb{R}^d} \sum_{i=1}^{n} \Phi_p(\boldsymbol{u}, \boldsymbol{X}_i - Q). \tag{5.2}$$

Observe at this point that a vector $\boldsymbol{u}$ for which $\|\boldsymbol{u}\|_q$ is close to one corresponds to an extreme quantile whereas a vector $\boldsymbol{u}$ for which $\|\boldsymbol{u}\|_q$ is close to zero corresponds to a central quantile. Since the vector $\boldsymbol{u}$ has a direction in addition to its magnitude, this immediately leads to a notion of directional outlyingness of a point with respect to the center of a cloud of observations based on the geometry of the cloud. It is also noteworthy that if we view the d-dimensional space $\mathbb{R}^d$ equipped with l_q-norm as the dual of the Banach space $\mathbb{R}^d$ equipped with l_p-norm where $1/p + 1/q = 1$, our index vector $\boldsymbol{u}$ is an element of the open unit ball in that dual space.

It is easy to observe that for $1 \leq p < \infty$, l_p-quantiles are not equivariant under arbitrary affine transformations of the data vectors and they are not even equivariant under orthogonal transformations unless $p = 2$ (for rotational equivariance in the case $p = 2$ see Chaudhuri 1996). Thus when the coordinate variables are measured in different units, or they have different degrees of statistical variation l_p-quantiles do not make much sense. This lack of affine equivariance makes l_p-quantiles very much dependent on the choice of the coordinate system, which is not at all desirable.

Let us now consider n data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$, and assume that $n > d+1$. Let $\alpha = \{i_0, i_1, \ldots, i_d\}$ denote a subset of size $(d+1)$ of $\{1, 2, \ldots, n\}$. Consider the points $\boldsymbol{X}_{i_0}, \boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_d}$ which will form a 'data-driven coordinate system', where $\boldsymbol{X}_{i_0}$ will determine the origin and the lines joining that origin to the remaining d data points $\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_d}$ will form various coordinate axes. The $d \times d$ matrix $\mathbf{X}(\alpha)$ containing the columns $\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{i_d} - \boldsymbol{X}_{i_0}$ can be taken as the transformation matrix for transforming the remaining data points $\boldsymbol{X}_j$'s $1 \le j \le n$, $j \notin \alpha$ to express them in terms of the new coordinate system as $\boldsymbol{Y}_j^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_j$. If the $\boldsymbol{X}_j$'s are i.i.d. observations with a common probability distribution that happens to be absolutely continuous w.r.t the Lebesgue measure on $\mathbb{R}^d$, $\mathbf{X}(\alpha)$ must be an invertible matrix with probability one. To compute the u-th l_p-quantile for $1 \le p < \infty$ and $\|u\|_q < 1$ with $1/p + 1/q = 1$ define

$$\begin{aligned} v(\alpha) &= \frac{\{\mathbf{X}(\alpha)\}^{-1}u}{\|\{\mathbf{X}(\alpha)\}^{-1}u\|_q}\|u\|_q \quad \text{for } u \neq 0 \\ &= 0 \qquad\qquad\qquad\quad \text{for } u = 0 \end{aligned}$$

Let $\hat{R}_n^{(\alpha,p)}(u)$ be the $v(\alpha)$-th l_p-quantile based on $\boldsymbol{Y}_j^{(\alpha)}$'s with $1 \le j \le n$, $j \notin \alpha$ as defined in (5.2). Then define the multivariate transformation retransformation (TR) l_p-quantile $\hat{Q}_n^{(\alpha,p)}(u)$ for the original data by retransforming $\hat{R}_n^{(\alpha,p)}(u)$ as $\hat{Q}_n^{(\alpha,p)}(u) = \{\mathbf{X}(\alpha)\}\hat{R}_n^{(\alpha,p)}(u)$. Note that as we transform the observations in the new coordinate system, we need to suitably modify the orientation of the index vector u. In the new coordinate system, the vector u should be transformed to $\{\mathbf{X}(\alpha)\}^{-1}u$, but it may not be in the open unit ball $B_q^{(d)}$. To preserve the l_q-norm of the vector u, we rescale $\{\mathbf{X}(\alpha)\}^{-1}u$ by multiplying it with $\|u\|_q/\|\{\mathbf{X}(\alpha)\}^{-1}u\|_q$. In the transformed coordinate system, we compute $v(\alpha)$-th l_p-quantile based on transformed observations and then retransform it back to the original coordinate system. We now state a Theorem demonstrating the equivariance of the TR l_p-quantile under arbitrary affine transformations of data vectors.

Theorem 5.2.1 *Let the d-dimensional random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be transformed to $\mathbf{A}\boldsymbol{X}_1 + b, \mathbf{A}\boldsymbol{X}_2 + b, \ldots, \mathbf{A}\boldsymbol{X}_n + b$, where $\mathbf{A}$ is a $d \times d$ nonsingular matrix and b is a vector in $\mathbb{R}^d$. Then for $w = (\|u\|_q/\|\mathbf{A}u\|_q)\mathbf{A}u$, the w-th TR l_p-quantile based on $\mathbf{A}\boldsymbol{X}_1 + b, \ldots, \mathbf{A}\boldsymbol{X}_n + b$ is given by $\mathbf{A}\hat{Q}_n^{(\alpha,p)}(u) + b$, where $\hat{Q}_n^{(\alpha,p)}(u)$ is the u-th TR l_p-quantile based on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$.*

Proof: As d-dimensional random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are transformed to $\mathbf{A}\boldsymbol{X}_1 + b, \ldots, \mathbf{A}\boldsymbol{X}_n + b$, where $\mathbf{A}$ is a $d \times d$ nonsingular matrix and b is $d \times 1$ vector, the transformation matrix $\mathbf{X}(\alpha)$ gets transformed to $\mathbf{A}\mathbf{X}(\alpha)$. For $u \in B_q^{(d)}$, define $w = (\|u\|_q/\|\mathbf{A}u\|_q)\mathbf{A}u$. Note that the index vector $v(\alpha)$ based on original observations and

corresponding to u is defined as $[\|u\|_q/\|\{\mathbf{X}(\alpha)\}^{-1}u\|_q]\{\mathbf{X}(\alpha)\}^{-1}u$ and that based on transformed observations and corresponding to w is given by

$$v^*(\alpha) = \frac{\{\mathbf{AX}(\alpha)\}^{-1}w}{\|\{\mathbf{AX}(\alpha)\}^{-1}w\|_q}\|w\|_q = \frac{\{\mathbf{AX}(\alpha)\}^{-1}\mathbf{A}u}{\|\{\mathbf{AX}(\alpha)\}^{-1}\mathbf{A}u\|_q}\|u\|_q = v(\alpha).$$

Also, note that the $\mathbf{Y}_i^{(\alpha)}$'s will be transformed to $\mathbf{Z}_i^{(\alpha)} = \mathbf{Y}_i^{(\alpha)} + \{\mathbf{AX}(\alpha)\}^{-1}b$, and the $v(\alpha)$-th l_p-quantile is equivariant under a location shift of the data points. Hence, the $v(\alpha)$-th l_p-quantile based on $Z_i^{(\alpha)}$'s is transformed to $\hat{R}_n^{(\alpha,p)}(w) = \hat{R}_n^{(\alpha,p)}(u) + \{\mathbf{AX}(\alpha)\}^{-1}b$. Consequently $\hat{Q}_n^{(\alpha,p)}(w)$, the w-th TR l_p-quantile based on transformed observations, which is defined as $\{\mathbf{AX}(\alpha)\}\hat{R}_n^{(\alpha,p)}(w)$, will be equal to $\mathbf{A}\hat{Q}_n^{(\alpha,p)}(u) + b$. Thus the w-th TR l_p-quantile based on transformed observations $(\mathbf{A}X_i + b)$ is $\mathbf{A}\hat{Q}_n^{(\alpha,p)}(u) + b$, where $\hat{Q}_n^{(\alpha,p)}(u)$ is the u-th TR l_p-quantile based on original X_i's. □

It is easy to see that, if we take $p = 2$ and $\mathbf{A}$ happens to be an orthogonal matrix, then $\mathbf{A}u$-th quantile based on $\mathbf{A}X_1 + b, \ldots, \mathbf{A}X_n + b$ will be given by $\mathbf{A}\hat{Q}_n^{(\alpha,2)}(u) + b$ where $\hat{Q}_n^{(\alpha,2)}(u)$ is the u-th transformation retransformation geometric quantile based on $X_1, X_2, \ldots, X_n$ (cf. Fact 2.2.1 of Chaudhuri 1996).

It should be noted that general M-quantiles defined by Koltchinskii (1997) are not affine equivariant in nature and we can employ this transformation retransformation strategy to general M-quantiles also to make them affine equivariant. But we have decided to restrict ourselves to l_p-quantiles here mainly because in many practical situations l_1-quantiles and spatial (or l_2) quantiles turn out to be adequate to explore different statistically important geometric aspects of a multivariate data cloud, some of which we will see later. The mathematical treatment of l_p-quantiles is not much different from those of l_2-quantiles, and for each $p \geq 1$, the l_p-norm leads to a notion of multidimensional symmetry and associated symmetric probability distributions will have contours that coincide with the balls defined by the l_p-norm. The existence and uniqueness of TR l_p-quantiles are given in the following Fact.

Fact 5.2.2 *Consider the d-dimensional observations $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ and $\alpha = \{i_0, i_1, \ldots, i_d\} \subset \{1, 2, \ldots, n\}$ such that the matrix $\mathbf{X}(\alpha)$ as defined earlier is invertible. Then the TR l_p-quantile $\hat{Q}_n^{(\alpha,p)}(u)$ exists for any given u in the open unit ball $B_q^{(d)}$, where $1/p + 1/q = 1$. Further, for $d \geq 2$ and $1 < p < \infty$, it will be unique if the X_i's, $i \notin \alpha$ are not all carried by a single straight line in $\mathbb{R}^d$.*

Efficient algorithms for computing spatial median have been extensively studied by Gower (1974) and Bedall and Zimmermann (1979). Chaudhuri (1996) suggested an algorithm to compute geometric quantiles which is a minor modification of Newton-Raphson algorithm

for finding roots of multivariate equations. We now state a fact characterizing TR l_p-quantiles in terms of data points from which it is computed.

Fact 5.2.3 *Consider $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$ and $\alpha = \{i_0, i_1, \ldots, i_d\} \subset \{1, 2, \ldots, n\}$ such that the matrix $\mathbf{X}(\alpha)$, as defined earlier, is invertible, and $\hat{Q}_n^{(\alpha,p)}(u)$ is the u-th TR l_p-quantile based on these observations. If $\hat{Q}_n^{(\alpha,p)}(u) \neq \boldsymbol{X}_i$ for all $i \notin \alpha$, we have for $1 < p < \infty$ and $1/p + 1/q = 1$*

$$\sum_{i \notin \alpha} \frac{\nu[\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \hat{Q}_n^{(\alpha,p)}(u))]}{\|\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \hat{Q}_n^{(\alpha,p)}(u))\|_p^{p-1}} + (n-d-1)v(\alpha) = 0 \tag{5.3}$$

On the other hand, if $\hat{Q}_n^{(\alpha,p)}(u) = \boldsymbol{X}_i$ for some $i \notin \alpha$, we will have

$$\left\| \sum_{i \notin \alpha, \boldsymbol{X}_i \neq \hat{Q}_n^{(\alpha,p)}} \left\{ \frac{\nu[\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \hat{Q}_n^{(\alpha,p)}(u))]}{\|\{\mathbf{X}(\alpha)\}^{-1}(\boldsymbol{X}_i - \hat{Q}_n^{(\alpha,p)}(u))\|_p^{p-1}} + v(\alpha) \right\} \right\|_q$$

$$\leq (1 + \|v(\alpha)\|_q)[\#\{i : \boldsymbol{X}_i = \hat{Q}_n^{(\alpha,p)}(u)\}], \tag{5.4}$$

where $\nu[(x_1, x_2, \ldots, x_d)^T] = (sign(x_1)|x_1|^{p-1}, \ldots, sign(x_d)|x_d|^{p-1})^T$, $v(\alpha)$ as defined earlier, and # denotes the number of elements in a set.

This fact implies that one can use iterative methods like Newton-Raphson type method to compute $\hat{Q}_n^{(\alpha,p)}(u)$ for $1 < p < \infty$. For $p = 1$, l_1-quantiles are nothing but coordinatewise quantiles. Thus, after transformation, one has to compute coordinatewise quantiles of the transformed observations and then retransform it back. This shows the simplicity of the computation involved in TR l_p-quantiles once the transformation matrix is fixed. Both of Facts 5.2.2 and 5.2.3 follow from some minor modifications of some of the fundamental results in Kemperman (1987) and Chaudhuri (1996), and we will skip their proofs here.

5.3 Large Sample Properties: Main Results

Let us begin by introducing some notations. For any $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{u} \in B_q^{(d)}$, we will write $\varphi_p(\boldsymbol{u}, \boldsymbol{x}) = \nu(\boldsymbol{x})/\|\boldsymbol{x}\|_p^{p-1} + \boldsymbol{u}$ for $\boldsymbol{x} \neq 0$ and $\varphi_p(\boldsymbol{u}, 0) = \boldsymbol{u}$. Note that $\varphi_p(\boldsymbol{u}, \boldsymbol{x})$ is the gradient or first order derivative of the function $\Phi_p(\boldsymbol{u}, \boldsymbol{x})$ w.r.t $\boldsymbol{x}$ when $\boldsymbol{x} \neq 0$. Let $\Psi_p(\boldsymbol{x})$ denote the $d \times d$ Hessian matrix or the second order derivative of $\Phi_p(\boldsymbol{u}, \boldsymbol{x})$ for $1 < p < \infty$. So, for $\boldsymbol{x} \neq 0$,

$$\Psi_p(\boldsymbol{x}) = (p-1)\|\boldsymbol{x}\|_p^{1-p}[W_p(\boldsymbol{x}) - \frac{\nu(\boldsymbol{x})\{\nu(\boldsymbol{x})\}^T}{\|\boldsymbol{x}\|_p^p}],$$

where $W_p(\boldsymbol{x})$ is the diagonal matrix $diag(|x_1|^{p-2}, \ldots, |x_d|^{p-2})$. We will adopt the convention that $\Psi_p(0) = 0 =$ the zero matrix. Note that when $p = 1$, $\varphi_p(\boldsymbol{u}, \boldsymbol{x})$ becomes $(sign(x_1), \ldots, sign(x_d))^T + \boldsymbol{u}$ and $\Psi_p(\boldsymbol{x})$ is identically equal to 0.

5.3.1 Asymptotic Behavior of TR l_p-quantiles

Let us define $Q^{(\alpha,p)}(u)$ as

$$Q^{(\alpha,p)}(u) = \arg \min_{Q \in R^d} E^{(\alpha)}[\Phi_p(v(\alpha), \{\mathbf{X}(\alpha)\}^{-1}(X - Q)) - \Phi_p(v(\alpha), \{\mathbf{X}(\alpha)\}^{-1}X)]$$

where $E^{(\alpha)}$ denotes the conditional expectation given the X_i's for which $i \in \alpha$ and $v(\alpha)$ is as defined in Section 5.2. In this Section, the observations X_i's will be assumed to be i.i.d. observations with a common probability distribution having density $h(x)$ on $\mathbb{R}^d$. Let us define

$$D_1^{(\alpha,p)}(Q) = E^{(\alpha)}\{\Psi_p(\{\mathbf{X}(\alpha)\}^{-1}(X - Q))\},$$

and

$$D_2^{(\alpha,p)}(Q,u) = E^{(\alpha)}[\{\varphi_p(v(\alpha), \{\mathbf{X}(\alpha)\}^{-1}(X - Q))\}^T\{\varphi_p(v(\alpha), \{\mathbf{X}(\alpha)\}^{-1}(X - Q))\}].$$

Theorem 5.3.1 (Bahadur type representation of TR l_1-quantiles)
Assume that $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of i.i.d. observations with a common density $h(x)$. Fix $\alpha = \{i_0, i_1, \ldots, i_d\} \subset \{1, 2, \ldots, n\}$ and the matrix $\mathbf{X}(\alpha)$ and assume that the j-th marginal g_j of the density $f(y) = |\det\{\mathbf{X}(\alpha)\}|h\{\mathbf{X}(\alpha)y\}$ is differentiable and positive at $Q_j^{\#}(u)$, where $Q_j^{\#}(u)$ is the j-th element of $\{\mathbf{X}(\alpha)\}^{-1}Q^{(\alpha,1)}(u)$ for $i = 1, \ldots, d$. Then for any $u \in \mathbb{R}^d$ such that $\|u\|_\infty < 1$, and given the X_i's with $i \in \alpha$, we have

$$\hat{Q}_n^{(\alpha,1)}(u) - Q^{(\alpha,1)}(u) \quad =$$

$$n^{-1}\mathbf{X}(\alpha)\{D_f(\alpha)\}^{-1}\sum_{i \notin \alpha}\left\{Sign[\{\mathbf{X}(\alpha)\}^{-1}\{X_i - Q^{(\alpha,1)}(u)\}] + v(\alpha)\right\} + R_n(u), \qquad (5.5)$$

where $D_f(\alpha)$ is the diagonal matrix $diag(2g_1\{Q_1^{\#}(u)\}, \ldots, 2g_d\{Q_d^{\#}(u)\})$, Sign denotes the vector of coordinatewise signs, and as $n \to \infty$, the remainder term $R_n(u)$ is almost surely $O(n^{-3/4}(\log n)^{3/4})$.

Before we prove Theorem 5.3.1, we state an asymptotic representation of $\hat{Q}_n^{(1)}(u)$, which is the non-equivariant vector of coordinatewise sample quantiles. Consider $Q^{(1)}(u)$ as the vector of marginal quantiles of the population distribution function F.

Lemma 5.3.2 *Let the j-th marginal distribution F_j of F be twice differentiable and $f_j(Q_j^{(1)}(u)) > 0$ where f_j is the j-th marginal density for $1 \le j \le d$ and $Q^{(1)}(u) = (Q_1^{(1)}(u), \ldots, Q_d^{(1)}(u))^T$. Then*

$$\hat{Q}_n^{(1)}(u) - Q^{(1)}(u) = n^{-1}D_f^{-1}\sum_{i=1}^{n}[Sign(X_i - Q^{(1)}(u)) + u] + R_n(u), \qquad (5.6)$$

where D_f is the diagonal matrix $diag(2f_1(Q_1^{(1)}(u)), \ldots, 2f_d(Q_d^{(1)}(u)))$ and as $n \to \infty$, the remainder term $R_n(u)$ is almost surely $O(n^{-3/4}(\log n)^{3/4})$.

The above Lemma follows almost directly from Bahadur (1966) representation for sample quantiles in the univariate case and thus we omit the proof of this Lemma. For the vector of marginal quantiles, the asymptotic normality can be derived with weaker conditions and it is studied in detail by Babu and Rao (1988).

Proof of Theorem 5.3.1: Define $\boldsymbol{Z}_i^{(\alpha)}$ as $\boldsymbol{Z}_i^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}\boldsymbol{X}_i$. Then, given the $\boldsymbol{X}_i$'s for which $i \in \alpha$, the transformed observations $\boldsymbol{Z}_i^{(\alpha)}$'s with $i \notin \alpha$ are conditionally i.i.d. random vectors with common density $|\det\{\mathbf{X}(\alpha)\}|h\{\mathbf{X}(\alpha)z\}$. The conditions of Theorem 5.3.1 imply that the conditions of Lemma 5.3.2 hold for the density of transformed data. Thus using Lemma 5.3.2 for the coordinatewise quantiles of transformed observations $\boldsymbol{Z}_i^{(\alpha)}$'s for $1 \le i \le n$, $i \notin \alpha$, we have the representation in (5.5) for the TR l_1-quantile $\hat{\boldsymbol{Q}}_n^{(\alpha,1)}(u)$. □

Theorem 5.3.3 (Bahadur type representation of TR l_p-quantiles, $1 < p < \infty$) *Assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n, \ldots$ is a sequence of i.i.d. observations with a common density $h(x)$ which is bounded on every compact subset of $\mathbb{R}^d$ with $d \ge 2$. Fix $\alpha = \{i_0, i_1, \ldots, i_d\} \subset \{1, 2, \ldots, n\}$ and the matrix $\mathbf{X}(\alpha)$. Then for any fixed $u \in B_q^{(d)}$, where $1 < p < \infty$ and $1/p + 1/q = 1$, the expectation defining the matrix $D_1^{(\alpha,p)}[\boldsymbol{Q}^{(\alpha,p)}(u)]$ will exist as a finite and invertible matrix, and given the $\boldsymbol{X}_i$'s with $i \in \alpha$, we have*

$$\hat{\boldsymbol{Q}}_n^{(\alpha,p)}(u) - \boldsymbol{Q}^{(\alpha,p)}(u) =$$

$$n^{-1}\mathbf{X}(\alpha)[D_1^{(\alpha,p)}(\boldsymbol{Q}^{(\alpha,p)}(u))]^{-1}\sum_{i\notin\alpha}\varphi_p[v(\alpha), \{\mathbf{X}(\alpha)\}^{-1}\{\boldsymbol{X}_i - \boldsymbol{Q}^{(\alpha,p)}(u)\}] + R_n(u), \quad (5.7)$$

where as $n \to \infty$, $R_n(u)$ is almost surely $O(\log n/n)$ if $d \ge 3$, and when $d = 2$, $R_n(u)$ is almost surely $o(n^{-\beta})$ for any fixed β such that $0 < \beta < 1$.

Before we prove Theorem 5.3.3, let us prove a Lemma on asymptotic representation of non-equivariant l_p-quantiles $\hat{\boldsymbol{Q}}_n^{(p)}(u)$ for $1 < p < \infty$. Let $\boldsymbol{Q}^{(p)}(u)$ be the population l_p quantile and define the matrices $D_1^{(p)}(\boldsymbol{Q}) = E\{\Psi_p(\boldsymbol{X} - \boldsymbol{Q})\}$ and $D_2^{(p)}(\boldsymbol{Q}, u) = E\{[\varphi_p(u, \boldsymbol{X} - \boldsymbol{Q})][\varphi_p(u, \boldsymbol{X} - \boldsymbol{Q})]^T\}$. Note that, $D_1^{(p)}(\boldsymbol{Q})$ will be positive definite unless the distribution of $\boldsymbol{X}$ is completely supported on a straight line in $\mathbb{R}^d$, and the expectation defining $D_1^{(p)}(\boldsymbol{Q})$ will exist finitely for $d \ge 2$ whenever $\boldsymbol{X}$ has a density that is bounded on compact subsets of $\mathbb{R}^d$. These facts can be verified directly.

Lemma 5.3.4 *Assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n, \ldots$ is a sequence of independent and identically distributed random vectors in $\mathbb{R}^d$ such that their common density is bounded on every bounded subset of $\mathbb{R}^d$. Then for any fixed $u \in B_q^{(d)}$, where $1 < p < \infty$ and $1/p + 1/q = 1$,*

we have the following Bahadur type representation for the u-th l_p-quantile:

$$\hat{Q}_n^{(p)}(u) - Q^{(p)}(u) = n^{-1}[D_1^{(p)}(Q(u))]^{-1} \sum_{i=1}^{n} \varphi_p(u, X_i - Q^{(p)}(u)) + R_n(u), \qquad (5.8)$$

where as $n \to \infty$, $R_n(u)$ is almost surely $O(logn/n)$ if $d \geq 3$ and when $d = 2$, $R_n(u)$ is almost surely $o(n^{-\beta})$ for any fixed β such that $0 < \beta < 1$.

Proof: We present the proof of the lemma following arguments similar to those used to prove the main results in Chaudhuri (1992a, 1996) with suitable modifications. We split the proof in several parts to expose the key ideas. Koltchinskii (1997) obtained a similar representation theorem but with slower rate of convergence for the remainder term $R_n(u)$. It follows form his result that there exists a constant $K_1 > 0$ such that we have almost surely $\|\hat{Q}_n^{(p)}(u) - Q^{(p)}(u)\|_p \leq K_1$ for all n sufficiently large.

Now observe that $\sum_{i=1}^{n} \varphi_p(u, X_i - \hat{Q}_n^{(p)}(u))$ is bounded [cf. Kemperman (1987), Chaudhuri(1996)] with the convention $\varphi_p(u, 0) = u$. Consequently, an easy extension of Proposition 5.6 of Chaudhuri (1992a) implies the existence of a constant $K_2 > 0$ such that almost surely $\|\hat{Q}_n^{(p)}(u) - Q^{(p)}(u)\|_p \leq K_2 n^{-1/2}(\log n)^{1/2}$ for all n sufficiently large. Recall here that $Q^{(p)}(u)$ satisfies $E[\varphi_p(u, Q^{(p)}(u))] = 0$, and lemmas 5.3 and 5.4 of Chaudhuri (1992a) can be suitably modified to imply that the magnitude of the d-dimensional vector $\sum_{i=1}^{n} \varphi_p(u, X_i - Q)$ will explode to infinity almost surely as $n \to \infty$, unless Q lies inside a ball in $\mathbb{R}^d$ with center at $Q^{(p)}(u)$ and radius of the order $O(n^{-1/2}[\log n]^{1/2})$.

Let B_n be the subset of $\mathbb{R}^d$ defined as

$$B_n = \{(v_1, \ldots, v_d) | n^4 v_i = \text{ an integer and } |v_i| \leq K_2 n^{-1/2}(\log n)^{1/2} \text{ for } 1 \leq i \leq d\}.$$

For $Q \in \mathbb{R}^d$, define

$$\Delta(Q) = E\{\varphi_p(u, X_1 - Q)\} + \{D_1^{(p)}(Q^{(p)}(u))\}\{Q - Q^{(p)}(u)\},$$

and for $Q \in B_n$, define

$$\Lambda_n(Q^{(p)}(u), Q + Q^{(p)}(u))$$

$$= n^{-1} \sum_{i=1}^{n} \{\varphi_p(u, X_i - Q^{(p)}(u)) - \varphi_p(u, X_i - Q^{(p)}(u) - Q)\}$$

$$+ E\{\varphi_p(u, X_1 - Q^{(p)}(u) - Q)\}.$$

Consider a sample sequence $X_1, X_2, \ldots, X_n, \ldots$ such that, for all n sufficiently large, we have $\|\hat{Q}_n^{(p)}(u) - Q^{(p)}(u)\|_p \leq K_4(\log n/n)^{1/2}$, and $\|\hat{Q}_n^{(p)}(u) - Q_n^*\|_p \leq K_3(\log n/n)$ for some $K_3 > 0$ and Q_n^* is a point in $\mathbb{R}^d$ such that $Q_n^* - Q^{(p)}(u) \in B_n$, and Q_n^* is closest

to $\hat{Q}_n^{(p)}(u)$ in l_p-norm. If there are several choices for such a Q_n^*, we can choose any one of them. It is quite easy to verify (see the proof of Proposition 5.6 in Chaudhuri 1992a) that the collection of all sample sequences satisfying these requirements will form a set of probability one. Now, we can write

$$n^{-1}\sum_{i=1}^{n}\varphi_p(u, X_i - Q^{(p)}(u))$$

$$= \Lambda_n\{Q^{(p)}(u), Q_n^*\} + \frac{1}{n}\sum_{i=1}^{n}\varphi_p(u, X_i - Q_n^*) - \Delta(Q_n^*) + D_1^{(p)}(Q^{(p)}(u))\{Q_n^* - Q^{(p)}(u)\}.$$

Some minor modifications of the arguments used in the proof of Fact 5.8 and Lemma 5.9 in Chaudhuri (1992a) implies that for a fixed constant $M^* > 0$,

$$\sup_{\|Q-Q^{(p)}(u)\|_p \le M^*(\log n/n)^{1/2}} \|\Delta(Q)\|_q = O(\log n/n) \tag{5.9}$$

as $n \to \infty$, for $d \geq 3$. On the other hand for $d = 2$, we have

$$\sup_{\|Q-Q^{(p)}(u)\|_p \le M^*(\log n/n)^{1/2}} \|\Delta(Q)\|_q = o(n^{-\omega}) \tag{5.10}$$

as $n \to \infty$ for any constant ω such that $1/2 < \omega < 1$. We also have that for $d \geq 3$ there is a constant $K_5 > 0$ such that $\max_{Q\in B_n} \|\Lambda_n(Q^{(p)}(u), Q + Q^{(p)}(u))\|_q \leq K_5(\log n/n)$ almost surely for all n sufficiently large. Further, if $d = 2$, we have $\max_{Q\in B_n} \|\Lambda_n(Q^{(p)}(u), Q + Q^{(p)}(u))\|_q = o(n^{-\omega})$ almost surely as $n \to \infty$, where ω is any constant satisfying $0 < \omega < 1$. On the other hand, it is quite easy to verify (cf. the inequality (6) in the proof of proposition 5.6 in Chaudhuri 1992a) that $n^{-1}\sum_{i=1}^{n}\varphi_p(u, X_i - Q_n^*) = O(n^{-1}\log n)$ almost surely as $n \to \infty$.

The proof of the lemma is now complete using the positive definiteness of the matrix $D_1^{(p)}[Q^{(p)}(u)]$ together with the fact that $\|\hat{Q}_n^{(p)}(u) - Q_n^*\|_p$ is $O(n^{-4})$ as $n \to \infty$ along our chosen sample sequence. □

Proof of Theorem 5.3.3: Define $Y_i^{(\alpha)} = \{\mathbf{X}(\alpha)\}^{-1}X_i$, for $1 \leq i \leq n$, $i \notin \alpha$. Then, given the X_j's for which $j \in \alpha$, the transformed observations $Y_i^{(\alpha)}$'s with $i \notin \alpha$ are conditionally i.i.d random vectors with common density $|\det\{\mathbf{X}(\alpha)\}|h\{\mathbf{X}(\alpha)y\}$. As the density h is bounded on every bounded subset of $\mathbb{R}^d$, the conditions in Lemma 5.3.4 hold for transformed observations. Using Lemma 5.3.4 for representation of l_p-quantiles of transformed observations $Y_i^{(\alpha)}$'s, we have the representation in (5.7) for TR l_p-quantile $\hat{Q}_n^{(\alpha,p)}(u)$. □

It will be appropriate to note here that Chaudhuri (1996) established a Bahadur type representation of nonequivariant l_2-quantiles and Koltchinskii (1997) considered general

non-equivariant multivariate M-estimators and proved a Bahadur type linear representation for them with a slower convergence rate for the remainder term. The following Corollaries are easy consequence of the above two Theorems.

Corollary 5.3.5 *Under the assumptions of Theorem 5.3.1, for any fixed* $u \in \mathbb{R}^d$ *such that* $\|u\|_\infty < 1$, *the conditional distribution of* $\sqrt{n}\{\hat{Q}_n^{(\alpha,1)}(u) - Q^{(\alpha,1)}(u)\}$ *given the* X_i*'s with* $i \in \alpha$ *converges weakly to a d-dimensional normal distribution with zero mean and dispersion matrix*

$$\{\mathbf{X}(\alpha)\}\{D_f(\alpha)\}^{-1}[D_2^{(\alpha,1)}(Q^{(\alpha,1)}(u),u)]\{D_f(\alpha)\}^{-1}\{\mathbf{X}(\alpha)\}^T$$

as $n \to \infty$.

Corollary 5.3.6 *Under the assumptions of Theorem 5.3.3, for* $1 < p < \infty$ *and for any* $u \in B_q^{(d)}$ *where* $1/p + 1/q = 1$, *the conditional distribution of* $\sqrt{n}\{\hat{Q}_n^{(\alpha,p)}(u) - Q^{(\alpha,p)}(u)\}$ *given the* X_i*'s with* $i \in \alpha$ *converges weakly to a d-dimensional normal distribution with zero mean and dispersion matrix*

$$\{\mathbf{X}(\alpha)\}[D_1^{(\alpha,p)}\{Q^{(\alpha,p)}(u)\}]^{-1}[D_2^{(\alpha,p)}\{Q^{(\alpha,p)}(u),u\}][D_1^{(\alpha,p)}\{Q^{(\alpha,p)}(u)\}]^{-1}\{\mathbf{X}(\alpha)\}^T$$

as $n \to \infty$.

5.3.2 Selection of α

The asymptotic normal distribution of $\hat{Q}_n^{(\alpha,p)}(u)$ established in the preceding section and the form of the associated dispersion matrix clearly indicates that the performance of the TR l_p-quantiles will depend upon the choice of the transformation matrix $\mathbf{X}(\alpha)$. Hence it is important to select a suitable subset of indices α. Before we state any formal method for selecting the transformation matrix $\mathbf{X}(\alpha)$, let us consider the special cases of l_1 and l_2 TR medians discussed in Chapter 2, which will provide some valuable insights into the problem. Let us assume that $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed random variables with a common elliptically symmetric density $|\det(\Sigma)|^{-1/2} f\{(x-\theta)^T\Sigma^{-1}(x-\theta)\}$ where Σ is a $d \times d$ positive definite matrix, $\theta \in \mathbb{R}^d$ and $f(x^Tx)$ is a density in $\mathbb{R}^d$. The main message communicated by Theorem 2.2.2, 2.2.3 and 2.3.1 is that for $u = 0$ (i.e. in the case of multivariate median) and $p = 1$ or 2, we need to choose $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a matrix of the form $\lambda\mathbf{I}_d$, which is a diagonal matrix with all diagonal entries equal. In other words, the coordinate system represented by the matrix $\Sigma^{-1/2}\mathbf{X}(\alpha)$ should be as orthonormal as possible. It also implies that when $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ is chosen to be close to a diagonal matrix with all diagonal entries equal, the asymptotic efficiency of the

estimate $\hat{Q}_n^{(\alpha,p)}(0)$ becomes close to that of the l_p-median under spherically symmetric models, and it will be more efficient than l_p median in elliptically symmetric models for $p = 1$ or 2.

Keeping in mind the fact that the above selection procedure provides "the most efficient transformation" for the multivariate median problem, we propose to select the transformation matrix $\mathbf{X}(\alpha)$ in such a way that $\{\mathbf{X}(\alpha)\}^T\Sigma^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a diagonal matrix with all diagonal entries equal. Here Σ is the scatter matrix associated with the underlying distribution of the X_i's which may not necessarily be elliptically symmetric. If the second moments of the underlying distribution exist, Σ can be taken to be the variance covariance matrix of that distribution. Since Σ will be an unknown parameter in practice, we have to estimate that from the data, and we will need an affine equivariant estimate (say $\hat{\Sigma}$). After obtaining $\hat{\Sigma}$, we will try to choose $\mathbf{X}(\alpha)$ in such a way that the eigen values of the positive definite matrix $\{\mathbf{X}(\alpha)\}^T\hat{\Sigma}^{-1}\mathbf{X}(\alpha)$ becomes as equal as possible. To achieve this, our strategy will be to minimize the ratio between the arithmetic mean and the geometric mean of the eigenvalues. Since the arithmetic mean and the geometric mean of the eigenvalues of a symmetric matrix can be obtained from its trace and the determinant respectively, we do not need to compute individual eigenvalues. Define now $\mathbf{X}^*(\alpha) = |\det(\mathbf{X}(\alpha))|^{-1/d}\mathbf{X}(\alpha)$ and $\hat{\Sigma}^* = \{\det(\hat{\Sigma})\}^{-1/d}\hat{\Sigma}$ where $\hat{\Sigma}$ is a positive definite matrix computed from the data. Note that, the absolute values of the determinants of the newly defined matrices $\mathbf{X}^*(\alpha)$ and $\hat{\Sigma}^*$ are both equal to 1, and the operation can be viewed as a way of normalizing matrices. Then to select the optimal α according to the above mentioned criteria, we only have to minimize the trace of $\{\mathbf{X}^*(\alpha)\}^T\hat{\Sigma}^{*-1}\mathbf{X}^*(\alpha)$. Suppose that for the subset of indices $\hat{\alpha}$, the trace of $\{\mathbf{X}^*(\alpha)\}^T\hat{\Sigma}^{*-1}\mathbf{X}^*(\alpha)$ is minimized.

Theorem 5.3.7 *Assume that, the random vectors $X_1, X_2, \ldots, X_n$ are independent and identically distributed with a common density $h(x)$ which satisfies*

$$\int_{\mathbf{R}^d}\{h(x)\}^{d+1}\,dx < \infty.$$

Further assume that $\hat{\Sigma}^*$ converges in probability to a positive definite matrix Σ^*. Then $\det(\Sigma^*) = 1$, *and trace*$[\{\mathbf{X}^*(\hat{\alpha})\}^T\Sigma^{*-1}\mathbf{X}^*(\hat{\alpha})]/d$ *converges to* 1 *in probability as* $n \to \infty$.

Before we prove Theorem 5.3.7, let us present some auxiliary results.

Lemma 5.3.8 *Assume that the observations $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed with a common density $h(x)$ such that $\int_{\mathbf{R}^d}\{h(x)\}^{d+1}\,dx < \infty$. Let Γ be a positive definite matrix with determinant equal to 1 and $\bar{\alpha}$ minimizes $t(\alpha) =$ trace$[\{\mathbf{X}^*(\alpha)\}^T\Gamma^{-1}\mathbf{X}^*(\alpha)]/d$. Then $t(\bar{\alpha})$ converges in probability to 1 as $n \longrightarrow \infty$.*

Proof: Let $\mathbf{A}$ be a $d \times d$ positive definite matrix such that $\Gamma = \mathbf{A}\mathbf{A}^T$. Consider $\alpha = \{1, 2, \ldots, d+1\}$. As the underlying distribution of the $\boldsymbol{X}_i$'s are independent and identically distributed with a common density h, the joint probability density function of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{d+1}$ can be written as $\prod_{i=1}^{d+1} h(x_i)$. Now we make the following transformation of variables:

$$\boldsymbol{Y}_1 = \mathbf{A}^{-1}(\boldsymbol{X}_2 - \boldsymbol{X}_1), \ldots, \boldsymbol{Y}_d = \mathbf{A}^{-1}(\boldsymbol{X}_{d+1} - \boldsymbol{X}_1), \boldsymbol{Y}_{d+1} = \mathbf{A}^{-1}\boldsymbol{X}_1.$$

Then the joint density of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{d+1}$ is given by

$$h(\mathbf{A}y_{d+1}) \prod_{i=1}^{d} h\{\mathbf{A}(y_i + y_{d+1})\}. \tag{5.11}$$

Therefore, the joint density of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$ at the origin in $\mathbb{R}^{d \times d}$ is

$$\int_{\mathbf{R}^d} \{h(\mathbf{A}y)\}^{d+1} \, dy,$$

which is finite and positive by the condition assumed in the statement of the Lemma. This condition further implies that the map

$$(y_1, y_2, \ldots, y_d) \mapsto \int_{\mathbf{R}^d} h(\mathbf{A}y) \prod_{i=1}^{d} h\{\mathbf{A}(y_i + y)\} \, dy \tag{5.12}$$

from $\mathbb{R}^{d \times d}$ to $\mathbb{R}$ is everywhere continuous. Therefore the joint density of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$ must remain bounded away from zero in a neighbourhood of $0 \in \mathbb{R}^{d \times d}$. Consequently the probability of the event that the columns of $\mathbf{A}\mathbf{X}(\alpha)$ will be nearly orthogonal and of nearly same length (and hence $\{\mathbf{X}^*(\alpha)\}^T \Gamma^{-1} \mathbf{X}^*(\alpha)$ will be very close to $\mathbf{I}_d$) is bounded away from zero. In other words, we have for any $\epsilon > 0$,

$$P[\|\{\mathbf{X}^*(\alpha)\}^T \Gamma^{-1} \mathbf{X}^*(\alpha) - \mathbf{I}_d\|_1 < \epsilon] = p_\epsilon > 0. \tag{5.13}$$

Let $\alpha_1, \alpha_2, \ldots, \alpha_{k_n}$ be disjoints subsets of $\{1, 2, \ldots, n\}$ each with size $d+1$ such that $k_n \to \infty$ as $n \to \infty$ [e.g. k_n may be equal to $n/(d+1)$]. Then
$P[|trace[\{\mathbf{X}^*(\bar{\alpha})\}^T \Gamma^{-1} \mathbf{X}^*(\bar{\alpha}) - \mathbf{I}_d]| > \epsilon]$

$$\begin{aligned} &\leq P[|trace[\{\mathbf{X}^*(\alpha_j)\}^T \Gamma^{-1} \mathbf{X}^*(\alpha_j) - \mathbf{I}_d]| > \epsilon, \text{ for } 1 \leq j \leq k_n\} \\ &\leq (1 - p_\epsilon)^{k_n} \longrightarrow 0 \text{ as } n \to \infty \end{aligned}$$

Hence, the result follows. □

Proof of Theorem 5.3.7: For $M > 1$, define $K_M^n = \{\alpha = \{i_0, i_1, \ldots, i_d\} : \quad t(\alpha) \equiv trace[\{\mathbf{X}^*(\alpha)\}^T \Sigma^{*-1} \mathbf{X}^*(\alpha)]/d \leq M\}$. Then it is easy to see that there exists some $M_1 > 0$

such that for any $\alpha \in K_M^n$, $trace[\mathbf{X}^*(\alpha)\{\mathbf{X}^*(\alpha)\}^T] \leq M_1$. Observe that, for any $\alpha \in K_M^n$

$$\begin{aligned} \|\{\mathbf{X}^*(\alpha)\}^T\hat{\Sigma}^{*-1}\mathbf{X}^*(\alpha) - \{\mathbf{X}^*(\alpha)\}^T\Sigma^{*-1}\mathbf{X}^*(\alpha)\|_2 &\leq \|\mathbf{X}^*(\alpha)\|_2^2\|\hat{\Sigma}^{*-1} - \Sigma^{*-1}\|_2 \\ &\leq M_1\|\hat{\Sigma}^{*-1} - \Sigma^{*-1}\|_2. \end{aligned}$$

Now, since $\hat{\Sigma}^* \xrightarrow{p} \Sigma^*$ as $n \to \infty$, where Σ^* is a positive definite matrix with determinant equal to 1, we have

$$\sup_{\alpha \in K_M^n} \|\{\mathbf{X}^*(\alpha)\}^T\hat{\Sigma}^{*-1}\mathbf{X}^*(\alpha) - \{\mathbf{X}^*(\alpha)\}^T\Sigma^{*-1}\mathbf{X}^*(\alpha)\|_2 \xrightarrow{p} 0 \tag{5.14}$$

as $n \to \infty$. Since $\bar{\alpha}$ minimizes $t(\alpha)$, by taking Σ^* as Γ in Lemma 5.3.8, we must have with large probability $\bar{\alpha} \in K_M^n$ for all sufficiently large n. Therefore using the continuity of the trace function of a matrix, we have

$$|\hat{t}(\bar{\alpha}) - t(\bar{\alpha})| \xrightarrow{p} 0, \tag{5.15}$$

where $\hat{t}(\alpha) = trace[\{\mathbf{X}^*(\alpha)\}^T\hat{\Sigma}^{*-1}\mathbf{X}^*(\alpha)]/d$. Thus, for all sufficiently large n, $\hat{t}(\bar{\alpha}) \leq M_2$ with large probability for some $M_2 > 0$. In other words, for $\hat{\alpha}$ which minimizes $\hat{t}(\alpha)$, we have $\hat{t}(\hat{\alpha}) \leq M_2$. Therefore $trace[\mathbf{X}^*(\hat{\alpha})\{\mathbf{X}^*(\hat{\alpha})\}^T]/d$ is also bounded in probability as $n \to \infty$. This in turn ensures that $|\hat{t}(\hat{\alpha}) - t(\hat{\alpha})| \xrightarrow{p} 0$ as $n \to \infty$.

Next, since $\hat{\alpha}$ minimizes $\hat{t}(\alpha)$ and $\bar{\alpha}$ minimizes $t(\alpha)$, it follows by some straightforward analysis that $|\hat{t}(\hat{\alpha}) - t(\hat{\alpha})| < \epsilon$ and $|\hat{t}(\bar{\alpha}) - t(\bar{\alpha})| < \epsilon$ will imply that $|\hat{t}(\hat{\alpha}) - t(\bar{\alpha})| < \epsilon$. Hence, we have

$$P[|\hat{t}(\hat{\alpha}) - t(\bar{\alpha})| > \epsilon] \leq P[|\hat{t}(\hat{\alpha}) - t(\hat{\alpha})| > \epsilon] + P[|\hat{t}(\bar{\alpha}) - t(\bar{\alpha})| > \epsilon],$$

and consequently $|\hat{t}(\hat{\alpha}) - t(\bar{\alpha})| \xrightarrow{p} 0$ as $n \to \infty$. Finally, since

$$|t(\hat{\alpha}) - t(\bar{\alpha})| \leq |t(\hat{\alpha}) - \hat{t}(\hat{\alpha})| + |\hat{t}(\hat{\alpha}) - t(\bar{\alpha})|,$$

it follows from Lemma 5.3.8 that $t(\hat{\alpha})$ converges in probability to 1. □

Clearly, the integrability condition imposed on h in Theorem 5.3.7 will hold if h happens to be a bounded density on $\mathbb{R}^d$. In the case of elliptic symmetry with $h(x) = \{\det(\Sigma)\}^{-1/2} f\{(x-\theta)^T\Sigma^{-1}(x-\theta)\}$, this condition translates into an integrability condition on f, which is again trivially satisfied for any bounded spherically symmetric density f on $\mathbb{R}^d$. It is interesting to note that if the second moments of the distribution of $\boldsymbol{X}_i$'s exist, $\hat{\Sigma}$ can taken to be the usual sample variance covariance matrix and $\hat{\Sigma}^*$ will converge in probability to Σ^* where Σ^* is the normalized version of the variance covariance matrix of the distribution. In the case of elliptically symmetric distributions, one can use any consistent affine equivariant estimate of the associated scale matrix Σ upto a scalar multiple.

As an alternative affine equivariant modification of spatial median, Isogai (1985) and Rao (1988) suggested spatial median based on observations transformed by the square root of the usual variance covariance matrix. While transforming the data points by the square root of the sample variance covariance matrix is a popular approach, the resulting coordinate system does not have any simple geometric interpretation. Further, such a transformation cannot lead to an affine equivariant modification of multivariate location estimates which are obtained by minimizing the general l_p distances for a p different from 2, (see Chapter 2), and the limitation of that approach is primarily due to the fact that there does not exist a way to extract an affine equivariant square root of the sample variance covariance matrix. On the other hand, observe that in a sense our selection procedure gives an "affine equivariant estimate" of the matrix $\Sigma^{1/2}$ which is further justified from our next result.

Theorem 5.3.9 *Under the conditions assumed in Theorem 5.3.7, the positive definite matrix* $\mathbf{X}^*(\hat{\alpha})\{\mathbf{X}^*(\hat{\alpha})\}^T$ *converges in probability to the matrix* Σ^* *as* $n \to \infty$.

Lemma 5.3.10 *Let* $\{\mathbf{A}_n\}$ *be a sequence of* $d \times d$ *random positive definite matrices such that* $\det(\mathbf{A}_n) = 1$ *for all* $n \geq 1$ *and* $trace(\mathbf{A}_n) \xrightarrow{p} d$ *as* $n \to \infty$. *Then* $\mathbf{A}_n \xrightarrow{p} \mathrm{I}_d$ *as* $n \to \infty$.

Proof : Let the eigenvalues of the positive definite matrix $\mathbf{A}_n$ be $\lambda_{1:n} \leq \lambda_{2:n} \leq \cdots \leq \lambda_{d:n}$. Then, if we show that $\lambda_{1:n} \xrightarrow{p} 1$ and $\lambda_{d:n} \xrightarrow{p} 1$ as $n \to \infty$, the proof of the Lemma will be complete. If possible, suppose that $\lambda_{d:n}$ does not converge in probability to 1 as $n \to \infty$. Then there exists some $\epsilon > 0$ and $\delta > 0$ such that for infinitely many values all $n \geq 1$, we will have

$$P[\lambda_{d:n} > 1 + \epsilon] > \delta.$$

Define $\mu_n = (\lambda_{1:n} + \cdots + \lambda_{d-1:n})/(d-1)$, i.e. the average of the eigenvalues excluding the maximum one. Then, as the product of all the eigenvalues is 1, we have by the A.M.–G.M. inequality $\mu_n \geq \lambda_{d:n}^{-1/(d-1)}$. Thus we have

$$trace(\mathbf{A}_n)/d = \frac{\lambda_{d:n} + (d-1)\mu_n}{d} \geq \frac{\lambda_{d:n} + (d-1)\lambda_{d:n}^{-1/(d-1)}}{d} > 1 + \epsilon_1$$

for some $\epsilon_1 > 0$ whenever $\lambda_{d:n} > 1 + \epsilon$. Here ϵ_1 depends on ϵ and d only. Therefore

$$P[trace(\mathbf{A}_n)/d > 1 + \epsilon_1] \geq P[\lambda_{d:n} > 1 + \epsilon] > \delta,$$

which contradicts the fact that $trace(\mathbf{A}_n)$ converges in probability to d as $n \to \infty$. Hence, we must have $\lambda_{d:n} \xrightarrow{p} 1$ as $n \to \infty$.

As the maximum eigenvalue $\lambda_{d:n}$ converges to 1 and the determinant of the matrix $\mathbf{A}_n$ is 1, all other eigenvalues including the minimum one must converge to 1 in probability as $n \to \infty$. □

Proof of Theorem 5.3.9: Theorem 5.3.7 implies that $trace[\{\mathbf{X}^*(\hat{\alpha})\}^T \Sigma^{*-1} \mathbf{X}^*(\hat{\alpha})]$ tends to d in probability as $n \to \infty$. Hence, $\|\mathbf{X}^*(\hat{\alpha})\|_2$ must remain bounded in probability as $n \to \infty$. Also, since $\det[\{\mathbf{X}^*(\hat{\alpha})\}^T \Sigma^{*-1} \mathbf{X}^*(\hat{\alpha})] = 1$, Theorem 5.3.7 and Lemma 5.3.10 imply that

$$\{\mathbf{X}^*(\hat{\alpha})\}^T \Sigma^{*-1} \mathbf{X}^*(\hat{\alpha}) \xrightarrow{p} \mathrm{I}_d$$

as $n \to \infty$. The proof is now complete by observing the fact

$$\|\mathbf{X}^*(\hat{\alpha})\{\mathbf{X}^*(\hat{\alpha})\}^T - \Sigma^*\|_2 \leq \|\mathbf{X}^*(\hat{\alpha})\|_2^2 \|\{\mathbf{X}^*(\hat{\alpha})\}^{-1} \Sigma^* \{[\mathbf{X}^*(\hat{\alpha})]^T\}^{-1} - \mathrm{I}_d\|_2.$$

□

Our results hold for any consistent and affine equivariant estimate of Σ (or Σ^*) and one can use robust estimates of scale as discussed by Davies (1987), which however are computationally quite intensive. Note that, this 'data-driven coordinate system' is a widely applicable tool for converting non-equivariant (or non-invariant) procedures into equivariant (or invariant) procedures, which is not limited to only l_p-quantiles. Besides, it has a very nice and intuitively meaningful geometric interpretation, and an attractive feature of this data-based transformation retransformation strategy is the clean and elegant mathematical theory associated with the approach.

5.4 Applications

5.4.1 Quantile Contour Plots

In the univariate set-up the quantiles uniquely determine the population distribution, and the sample quantiles provide a fair idea about the shape of the distribution. While exploring a multivariate data cloud, one may be interested to find out quantile contours, which join the quantiles for which the length of the index vector $\boldsymbol{u}$ is a constant, to get ideas about the shape of the underlying population distribution. Thus quantile contours can be described by the sets $\{\hat{Q}_n^{(\alpha,p)}(\boldsymbol{u}) : \|\boldsymbol{u}\|_q = r\}$ where $0 < r < 1$. For $r = 0$, it comprises of only one point – the TR l_p median. In principle, quantile contours can be constructed for any dimension $d \geq 2$, but for practical purposes, it is easier to visualize things only for bivariate data.

It is interesting to note that, for the optimal selection of the transformation matrix $\mathbf{X}(\alpha)$, the population quantile contours corresponding to $p = 2$ are nothing but the level

sets of the probability density function (or, probability density contours) when the underlying distribution is elliptically symmetric with density $\{\det(\Sigma)\}^{-1/2} f\{(\boldsymbol{x}-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\}$. The optimal selection of $\mathbf{X}(\alpha)$ provides an estimate of the matrix $\Sigma^{1/2}$ upto a scalar multiple and premultiplying the observations by $\{\mathbf{X}(\alpha)\}^{-1}$ makes the data spherical. As probability density contours characterize a distribution, the affine equivariant TR l_2-quantile contour plots can be used to measure the closeness of the data to a specific elliptically symmetric probability distribution. Even when the underlying probability distribution is not elliptically symmetric, Koltchinskii (1997) observed that the spatial quantile process uniquely determines the population distribution. Vector of coordinatewise quantiles determine the marginals of the joint multivariate distribution. However, marginals do not uniquely determine the joint distribution. Thus TR l_1-quantile contour plots cannot be used as a tool for measuring proximity to a multivariate distribution. Nevertheless, they can provide some insights into the geometry of the multivariate data cloud and help in identifying possible outliers.

To illustrate quantile contour plots, we simulated 100 observations from bivariate normal populations with zero means, unit standard deviations and varying correlation coefficients $\rho = 0.0, 0.5$ and 0.95. In Figure 5.1 (a), (b) and (c), we have plotted TR l_1-quantiles for $r = 0.1, 0.2, \ldots, 0.9$. To construct quantile contours, for each r, we have taken 32 values of $\boldsymbol{u}$ such that $\|\boldsymbol{u}\|_\infty = r$ and joined the corresponding quantiles. In Figures 5.1 (d), (e) and (f) we have similarly plotted TR l_2-quantiles for $r = 0.1, 0.2, \ldots, 0.9$. For each r, we have computed quantiles corresponding to $\boldsymbol{u} = (r\cos\theta, r\sin\theta)^T$ where $\theta = \pi k/16$, $k = 0, 1, \ldots, 31$ and joined them. We notice that as the TR quantiles are affine equivariant, quantile contours nicely capture the shift of the distribution from spherical symmetry to elliptical symmetry. The regions enclosed by quantile contours can be viewed as multivariate analogs of box and whisker plots used for univariate data.

Another interesting application of these quantile contours is in detecting outliers in the multivariate data. In multidimension, it is really difficult to detect the outliers. Here we suggest a simple procedure. We compute the quantile contour for some r close to 1 (the choice of r depends on the problem and the user's preference), and if a particular observation lies outside this contour, then we will call it an outlier. We demonstrate the methodology in a real data set. Reaven and Miller (1979) examined the relationship between chemical, subclinical and overt nonketotic diabetes in 145 non-obese adult subjects. The three primary variables used in the analysis are glucose intolerance, insulin response to oral glucose and insulin resistance. In addition, the relative weight and fasting plasma glucose were also measured for each individual in the study conducted at the Stanford Clinical Research Center (see Table 5.1). We have taken only 76 overt nonketotic diabetic patients and in Figure 5.2 we have shown the TR l_2-quantile contours by taking two vari-

Table 5.1: Measures of blood glucose and insulin levels of overt diabetic patients

Patient number	Fasting plasma glucose	Glucose area	Insulin area	SSPG	Patient number	Fasting plasma glucose	Glucose area	Insulin area	SSPG
1	80	356	124	55	39	106	396	128	80
2	97	289	117	76	40	98	277	222	186
3	105	319	143	105	41	102	378	165	117
4	90	356	199	108	42	90	360	282	160
5	90	323	240	143	43	94	291	94	71
6	86	381	157	165	44	80	269	121	29
7	100	350	221	119	45	93	318	73	42
8	85	301	186	105	46	86	328	106	56
9	97	379	142	98	47	85	334	118	122
10	97	296	131	94	48	96	356	112	73
11	91	353	221	53	49	88	291	157	122
12	87	306	178	66	50	87	360	292	128
13	78	290	136	142	51	94	313	200	233
14	90	371	200	93	52	93	306	220	132
15	86	312	208	68	53	86	319	144	138
16	80	393	202	102	54	86	349	109	83
17	90	364	152	76	55	96	332	151	109
18	99	359	185	37	56	86	323	158	96
19	85	296	116	60	57	89	323	73	52
20	90	345	123	50	58	83	351	81	42
21	90	378	136	47	59	100	398	122	176
22	88	304	134	50	60	110	426	117	118
23	95	347	184	91	61	80	333	131	136
24	90	327	192	124	62	96	418	130	153
25	92	386	279	74	63	95	391	137	248
26	74	365	228	235	64	82	390	375	273
27	98	365	145	158	65	84	416	146	80
28	100	352	172	140	66	100	385	192	180
29	86	325	179	145	67	86	393	115	85
30	98	321	222	99	68	93	376	195	106
31	70	360	134	90	69	107	403	267	254
32	99	336	143	105	70	112	414	281	119
33	75	352	169	32	71	93	364	156	159
34	90	353	263	165	72	93	391	221	103
35	85	373	174	78	73	90	356	199	59
36	99	376	134	80	74	99	398	76	108
37	100	367	182	54	75	93	393	490	259
38	78	335	241	175	76	89	318	73	220

ables at a time and $r = 0.0, 0.1, \ldots, 0.9$. These quantile contours clearly reveal that there are some outliers in the data set. Note that affine equivariance of the quantiles is crucial in outlier detection as the outlyingness of a data point should not be judged differently in different coordinate systems.

5.4.2 Multivariate Ranks

In univariate set-up, the concept of ranks and quantiles are closely related. Jan and Randles (1994) and Möttönen and Oja (1995) considered some notions of multivariate ranks which are closely related to geometric quantiles (or l_2-quantiles). Chaudhuri (1996) suggested the d-dimensional direction vector $n^{-1} \sum_{X_i \neq y} \|X_i - y\|_2^{-1}(X_i - y)$ as the multivariate rank of $y \in \mathbb{R}^d$. We may define affine invariant notions of multivariate ranks based on our transformation retransformation approach as follows. Consider the d-dimensional direction vector based on l_2-norm $n^{-1} \sum_{X_i \neq y, i \notin \alpha} \|\{\mathbf{X}(\alpha)\}^{-1}(X_i - y)\|_2^{-1}\{\mathbf{X}(\alpha)\}^{-1}(X_i - y)$ or alternatively based on l_1-norm $n^{-1} \sum_{X_i \neq y, i \notin \alpha} Sign[\{\mathbf{X}(\alpha)\}^{-1}(X_i - y)]$, which can be viewed as descriptive statistics that determine the geometric position of the point $y \in \mathbb{R}^d$ with respect to the data cloud formed by the observations $X_1, X_2, \ldots, X_n$, and these lead to vector valued concepts of multivariate centered ranks corresponding to TR l_2 and l_1-quantiles. Similarly, from the gradient vectors of the other l_p-norms, one can construct different versions of multivariate ranks. However, it is rather easy to interpret and geometrically visualize things for $p = 1$ and $p = 2$. Observe that the multivariate rank vectors associated with TR l_p-quantiles lie inside the unit ball $B_q^{(d)}$ where as usual $1/p + 1/q = 1$. There are some attempts to construct ranks as univariate quantities based on different data depth concepts like Tukey's half-space depth (Tukey, 1975) and Liu's simplicial depth (Liu, 1990), but they fail to take into account the orientation of a point in the data cloud. Univariate concepts of ranks can distinguish between 'extreme' points and 'central' points but they do not provide the information whether the 'extreme' observations are 'low' or 'high' observations with respect to some specific directions. For these limitations multivariate notions of ranks are often preferred over univariate notions. Based on these affine invariant multivariate ranks one can construct different rank related methodologies in multidimension extending univariate rank based methodologies.

5.4.3 Multivariate Q-Q Plots

Q-Q plots are popular and useful diagnostic tools in univariate data analysis. With their help, it is possible to assess graphically the closeness of a sample to a particular univariate

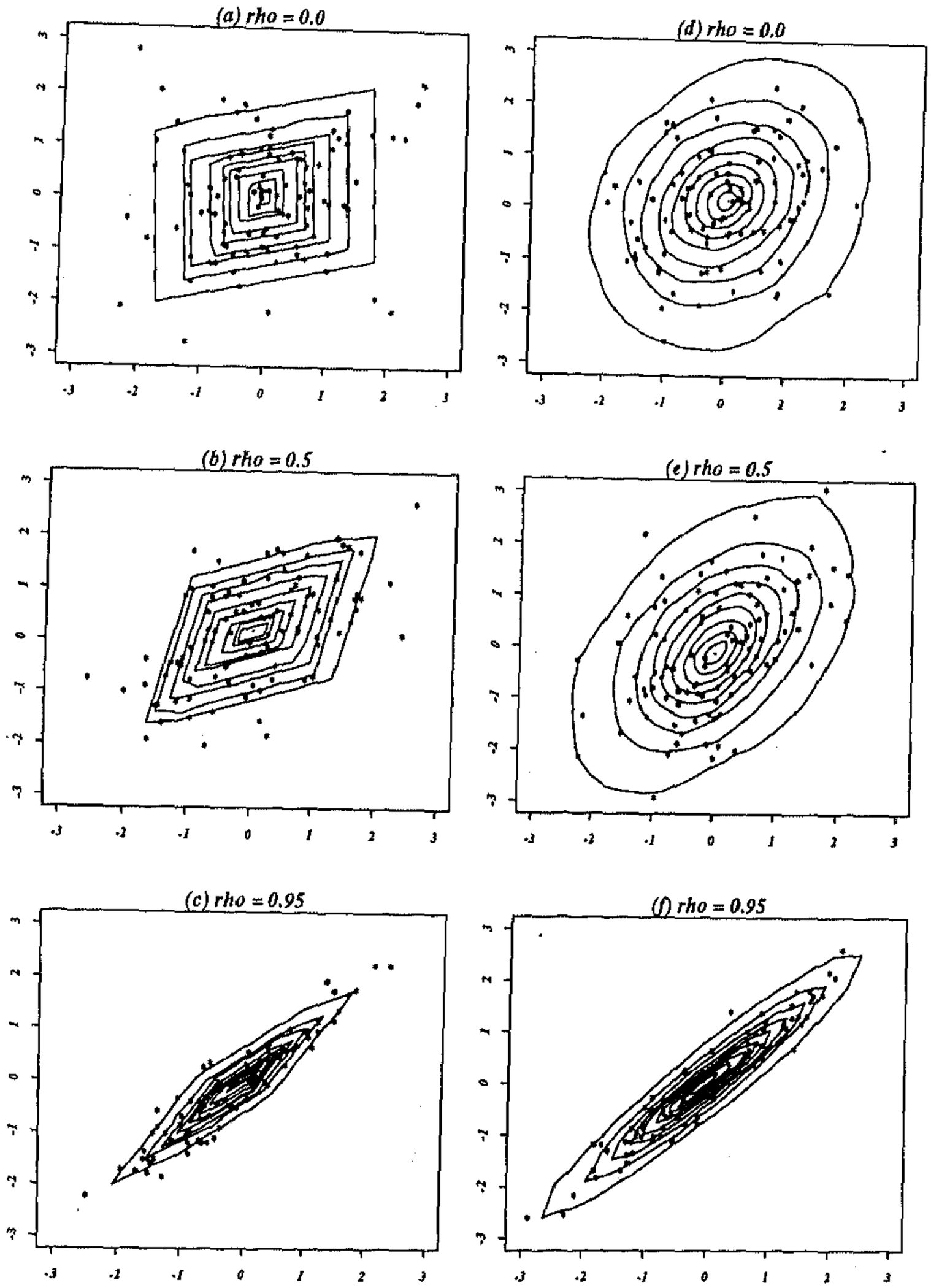
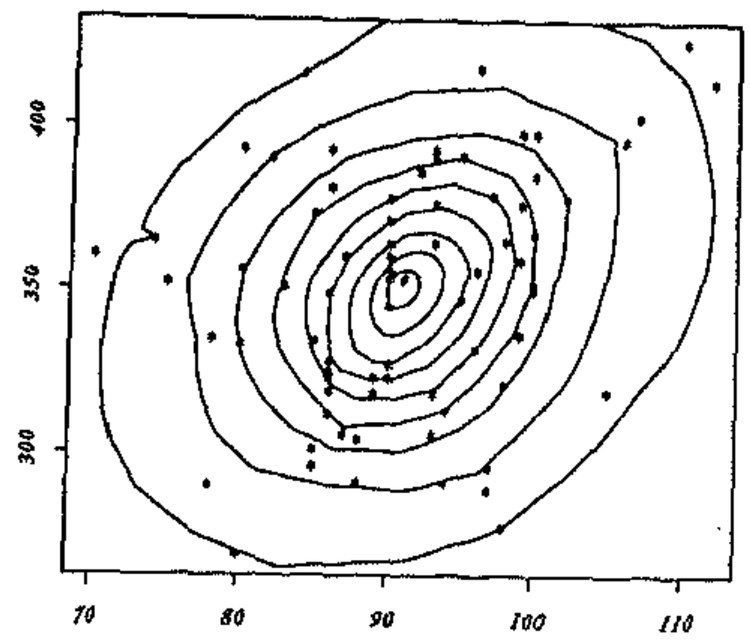
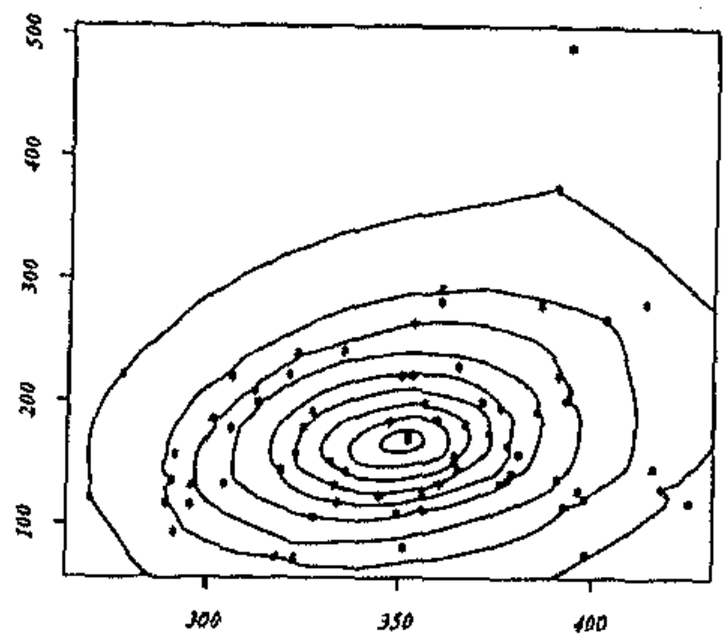


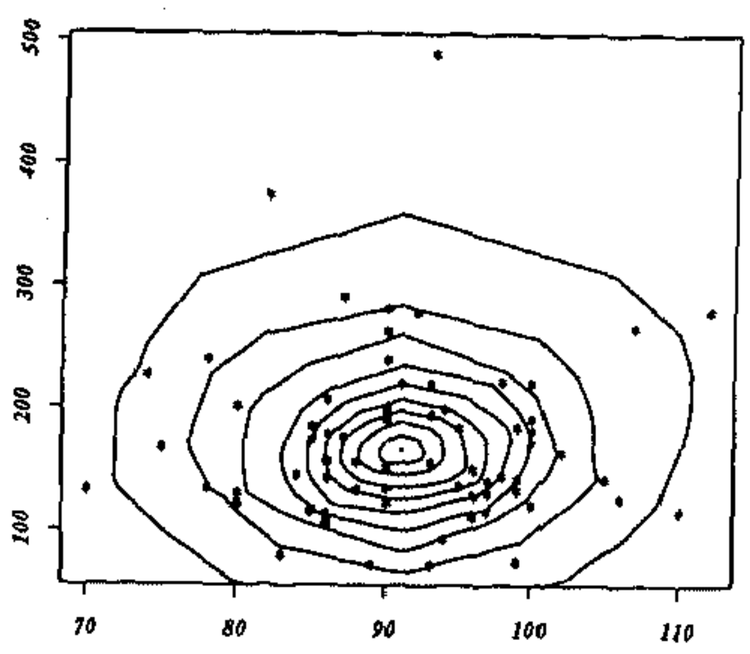
Figure 5.1: l_1 and l_2 quantile contour plots for bivariate normal data with different correlation coefficient



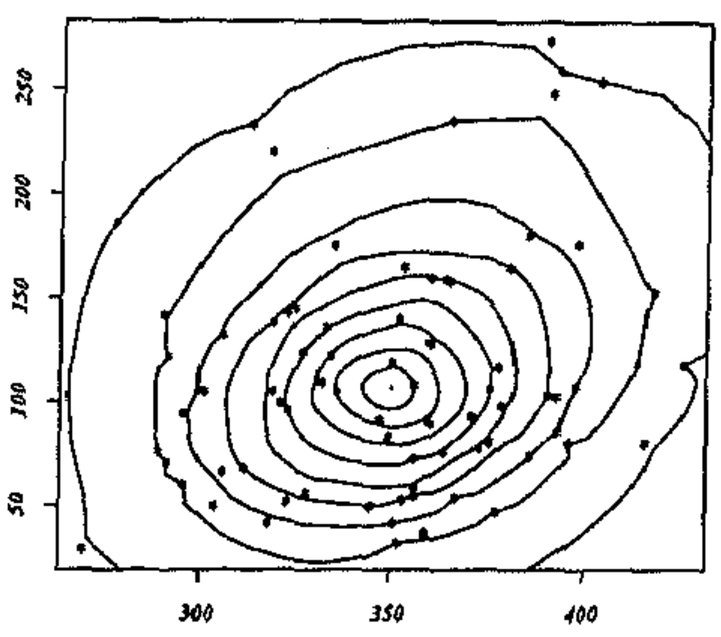
Fasting Glucose vs. Glucose Intolerance



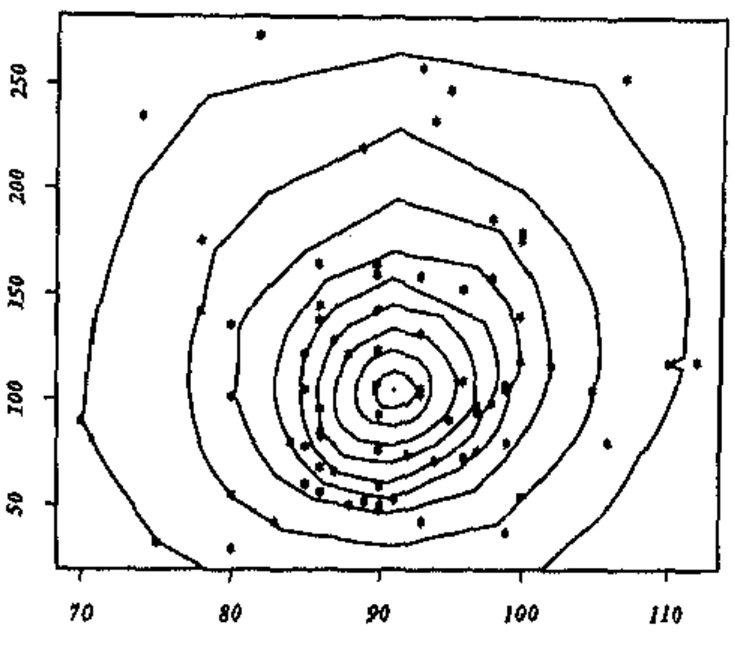
Glucose Resistance vs. Insulin Response



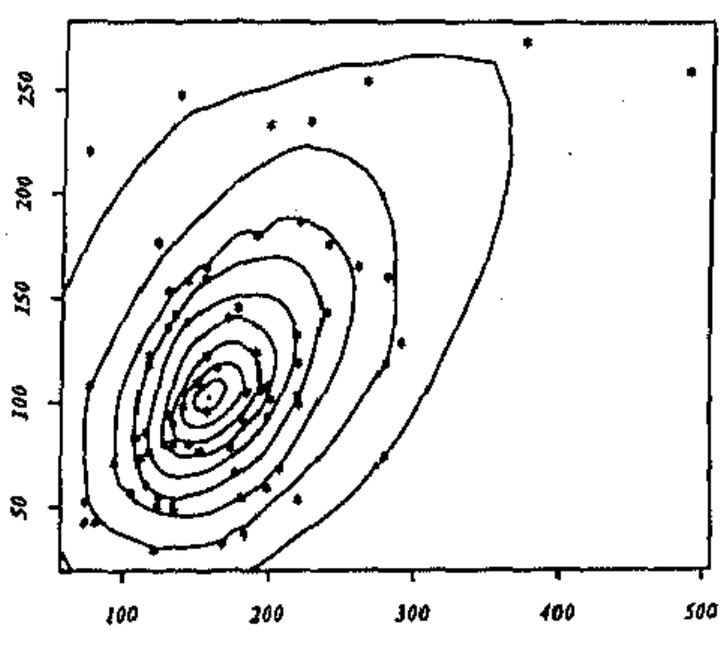
Fasting Glucose vs. Insulin Response



Glucose Resistance vs. Insulin Resistance



Fasting Glucose vs. Insulin Resistance



Insulin Response vs. Insulin Resistance

Figure 5.2: Quantile contour plots of blood sugar data for overt diabetic patients

distribution or the closeness between two independent samples. The idea behind Q-Q plots is to compute and plot a finite number of quantiles from the sample and corresponding quantiles from the comparing probability distribution or from the comparing sample. Marden (1998) generalized this concept to define bivariate Q-Q plots. He started by computing multivariate ranks associated with geometric quantiles for each observation in the sample, and then he computed corresponding geometric quantile from the comparing probability distribution by taking the rank of the observation as the index vector u. Then he joined the original data point and the quantile from the distribution with a directed arrow. If these arrows are very small in length and randomly oriented, then one may conclude that the sample does not deviate much from the chosen probability distribution. But if most of the arrows are directed towards a particular direction then the sample is more skewed in that direction, and if in general arrow lengths are large then the sample obviously does not conform with the given distribution. But these Q-Q plots are not affine invariant in nature and thus the presence of high correlations among the coordinate variables will often lead to inappropriate inference. To resolve the problem, we employ transformation retransformation technique. At first, one should transform the data points by $\{\hat{\lambda}\mathbf{X}(\alpha)\}^{-1}$ where $\hat{\lambda}^2 = trace[\{\mathbf{X}(\alpha)\}^{-1}\hat{\Sigma}(\{\mathbf{X}(\alpha)\}^T)^{-1}]/d$. After that l_p-ranks of the transformed observations are computed as discussed earlier. As we know that these l_p-rank vectors lie in $B_q^{(d)}$, and one can compute corresponding l_p-quantiles of the comparing probability distribution with scatter matrix $\mathbf{I}_d$ and location parameter $\mathbf{0}$. Then following Marden (1998), we should plot the arrows from the l_p-quantiles of the given distribution to the transformed observations. We have noted earlier that a proper selection of the transformation matrix $\mathbf{X}(\alpha)$ leads to an estimate of the scatter matrix Σ and $\hat{\lambda}^{-2}\{\mathbf{X}(\alpha)\}^{-1}\Sigma\{[\mathbf{X}(\alpha)]^T\}^{-1}$ is expected to be close to a d-dimensional identity matrix.

The TR Q-Q plots, which are affine invariant, can be used to construct tests of goodness of fit to a given multivariate distribution. There is no known good way of testing in practice whether the observed data is from a specified multivariate distribution or not. Both of the well-known χ^2-goodness of fit test and Kolmogorov-Smirnov test have serious practical limitations and are not very useful for multivariate problems. We suggest the following test procedure. At first, we make the data spherical by transforming the observations by $\{\hat{\lambda}\mathbf{X}(\alpha)\}^{-1}$ and then subtract the l_p-median of the transformed observations from them. Let us call these observations $\boldsymbol{Z}_i^{(\alpha,p)}$'s for $i \notin \alpha$. Thus, we have transformed observations with location parameter zero and identity as the scale matrix. Then we compute l_p-ranks of each of these transformed observations and corresponding l_p-quantiles of the population distribution (say, $Q_i^{(\alpha,p)}$'s). In the case $p = 2$, these population l_p-quantiles can be computed using the formula given in Möttönen, Oja and Tienari (1997) for spherically symmetric distributions. After that, let us consider the statistic $T_n^{(\alpha,p)} =$

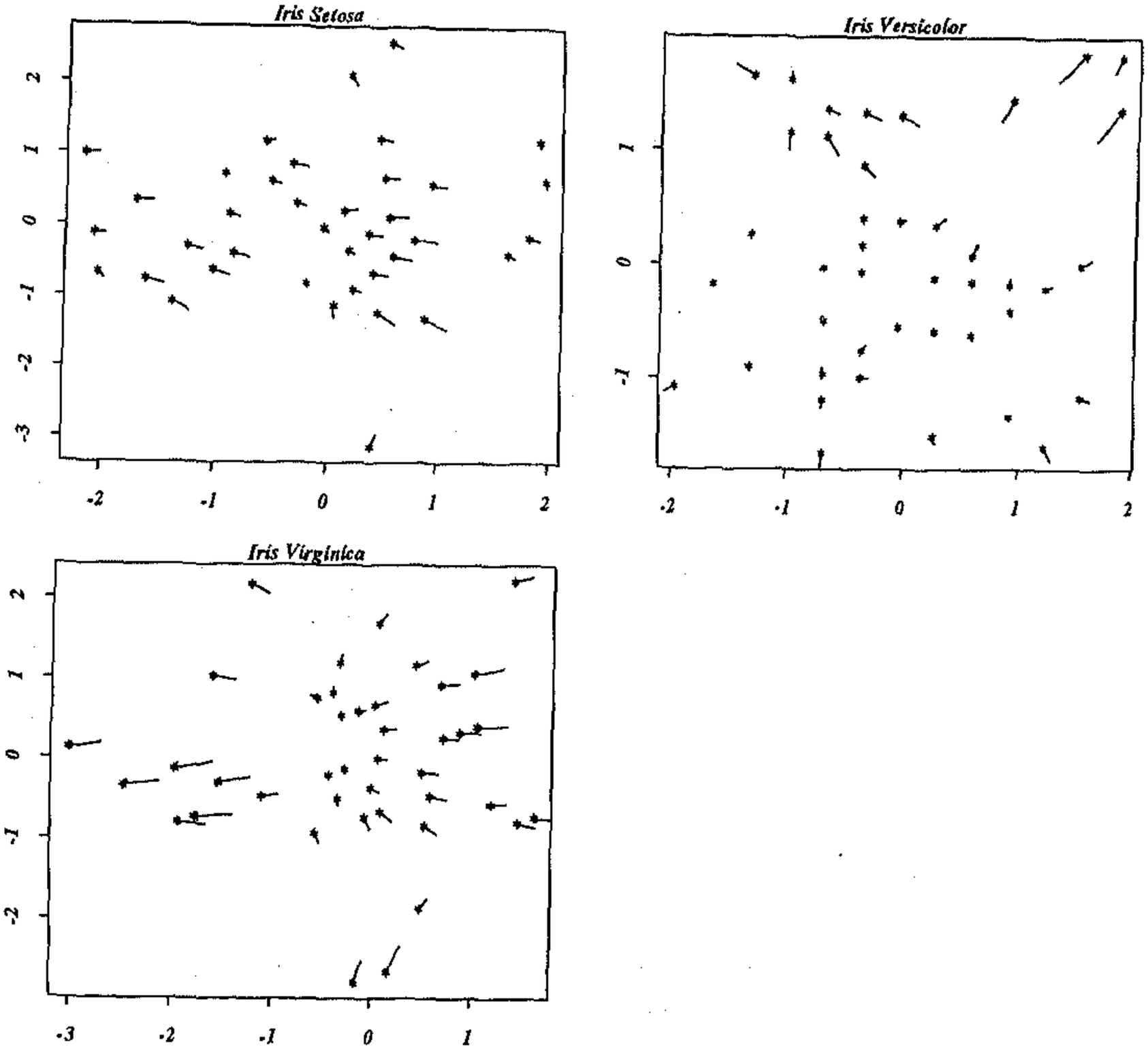


Figure 5.3: Multivariate Q-Q plot for Fisher's Iris data

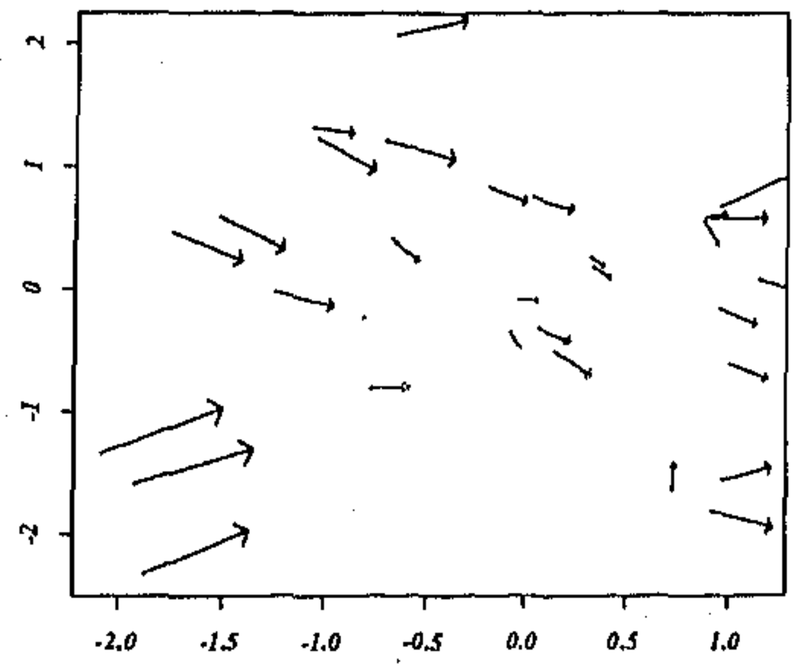
$\sum_{i \notin \alpha}(\boldsymbol{Z}_i^{(\alpha,p)} - \boldsymbol{Q}_i^{(\alpha,p)})$, which is nothing but the sum of the "directed arrows" as discussed earlier. If the observed data is close to the given distribution, the norm of the vector $T_n^{(\alpha,p)}$ should be close to zero. Thus $T_n^{(\alpha,p)}$ can be used as a test statistic for testing goodness of fit of a multivariate distribution. As an illustration, we have constructed bivariate affine invariant Q-Q plots for Iris Setosa, Iris Virginica and Iris Versicolor of the famous Fisher's iris data using TR l_2-quantiles. We have considered only two variables sepal length and sepal width for the demonstration purpose and compared them with bivariate normal distributions in Figure 5.3, where the plots indicate fairly good fits. Möttönen, Oja and Tienari (1997) provided a result for computing geometric quantiles of the bivariate normal distribution, and we have used that for our calculations.

As discussed earlier, we can also construct tests of equality of the underlying distri-

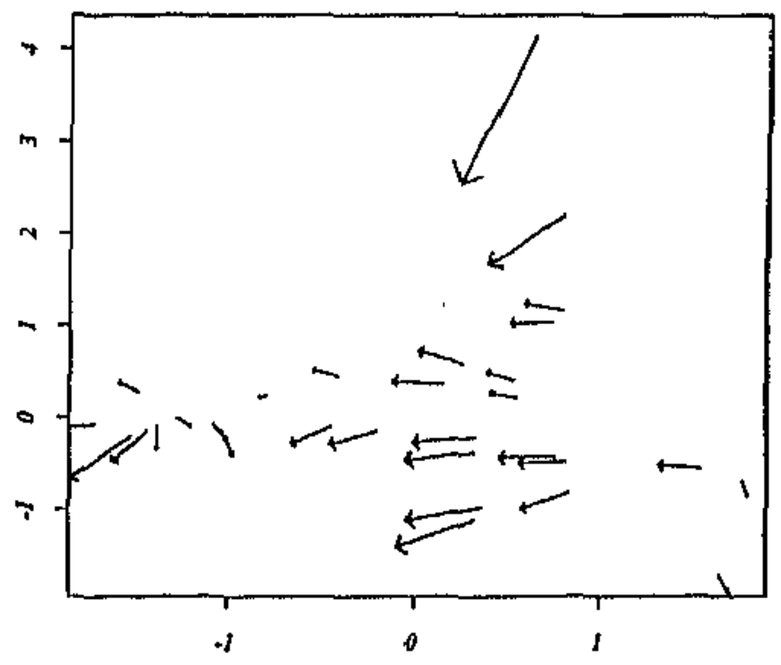
Table 5.2: Measures of blood glucose and insulin levels of normal patients

Patient number	Fasting plasma glucose	Glucose area	Insulin area	SSPG	Patient number	Fasting plasma glucose	Glucose area	Insulin area	SSPG
1	300	1468	28	455	18	146	847	103	339
2	303	1487	23	327	19	124	538	460	320
3	125	714	232	279	20	213	1001	42	297
4	280	1470	54	382	21	330	1520	13	303
5	216	1113	81	378	22	123	557	130	152
6	190	972	87	374	23	130	670	44	167
7	151	854	76	260	24	120	636	314	220
8	303	1364	42	346	25	138	741	219	209
9	173	832	102	319	26	188	958	100	351
10	203	967	138	351	27	339	1354	10	450
11	195	920	160	357	28	265	1263	83	413
12	140	613	131	248	29	353	1428	41	480
13	151	857	145	324	30	180	923	77	150
14	275	1373	45	300	31	213	1025	29	209
15	260	1133	118	300	32	328	1246	124	442
16	149	849	159	310	33	346	1568	15	253
17	233	1183	73	458					

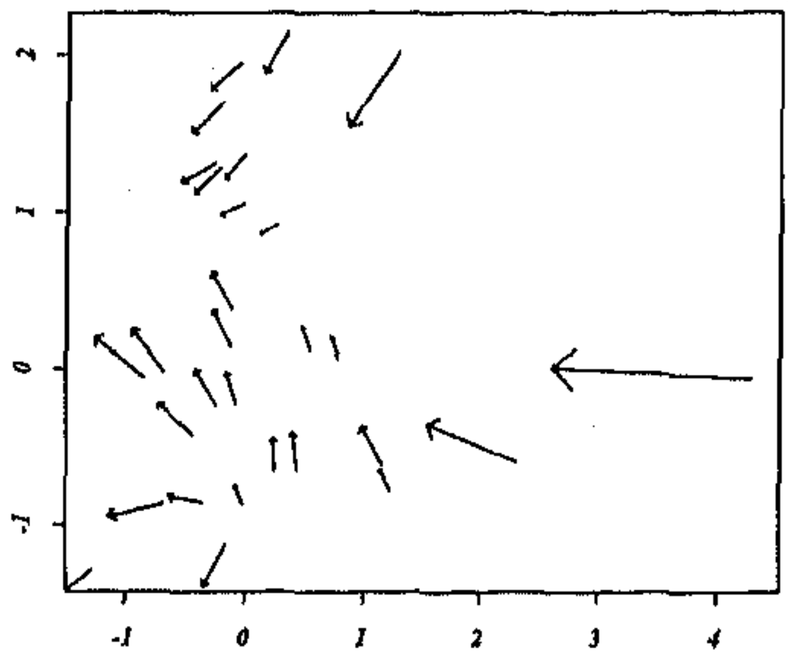
bution of two multivariate samples in a similar fashion. Here we compute transformed observations for both the samples and based on the ranks of the observations of one sample, we compute the quantiles of the other sample and draw directed arrows. Sum of these directed arrows provides us a test statistic for testing equality of the underlying distribution of two samples. To illustrate the comparison between two samples using Q-Q plots, we again used the blood sugar data, which we have used earlier to demonstrate quantile contour plots. In Figure 5.4, we construct Q-Q plots for comparing normal patients (see Table 5.2) *with overt nonketotic patients by computing multivariate affine invariant* l_2*-ranks* of the first sample and corresponding geometric quantiles of the transformed observations of the second sample. We have taken two variable at a time. From these Q-Q plots, it is quite apparent that the underlying distributions of the normal patients and overt diabetic patients are quite different. Large arrow lengths in all the plots suggest that there are possibly differences in locations and scales of the distributions and also the arrows are oriented towards a common direction indicating possible differences in the shapes of the distributions.



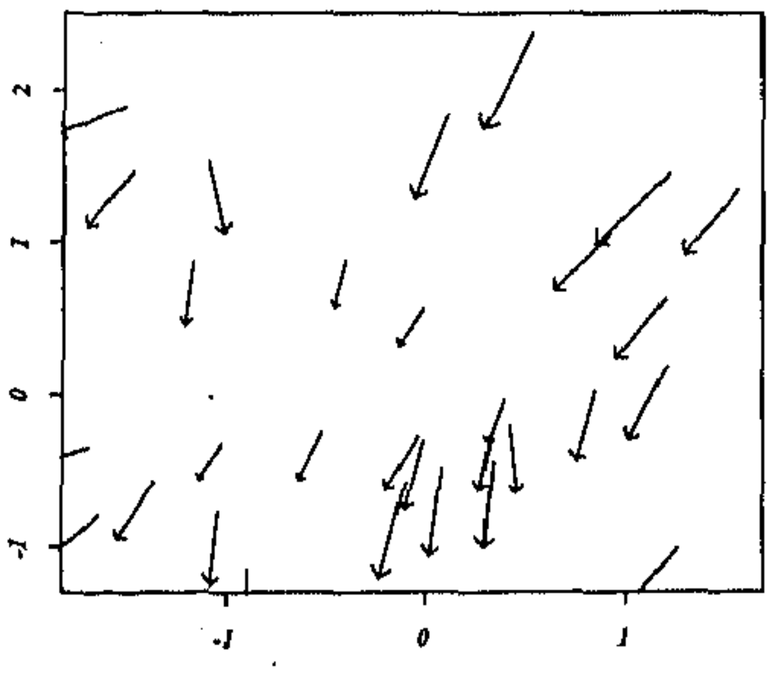
Fasting Glucose vs. Glucose Intolerance



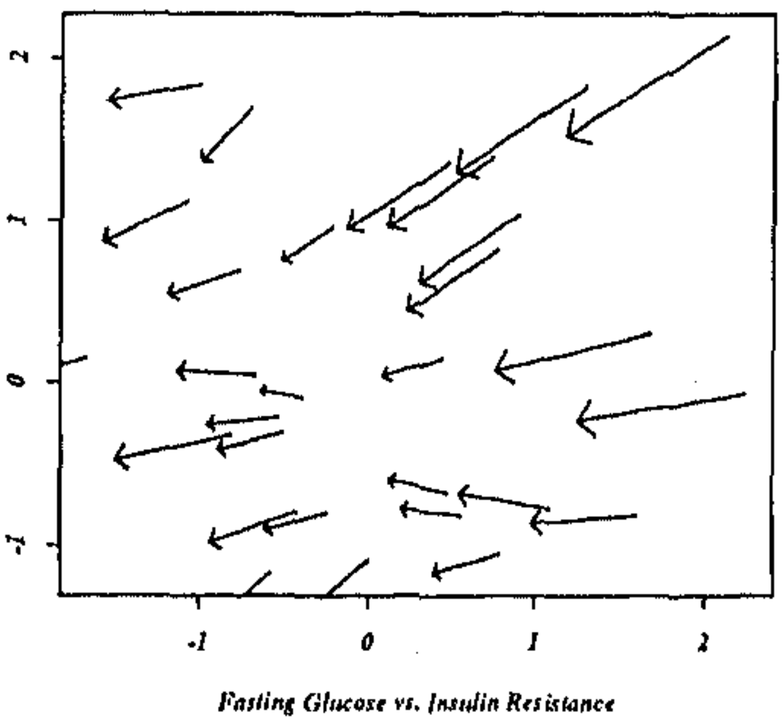
Glucose Resistance vs. Insulin Response



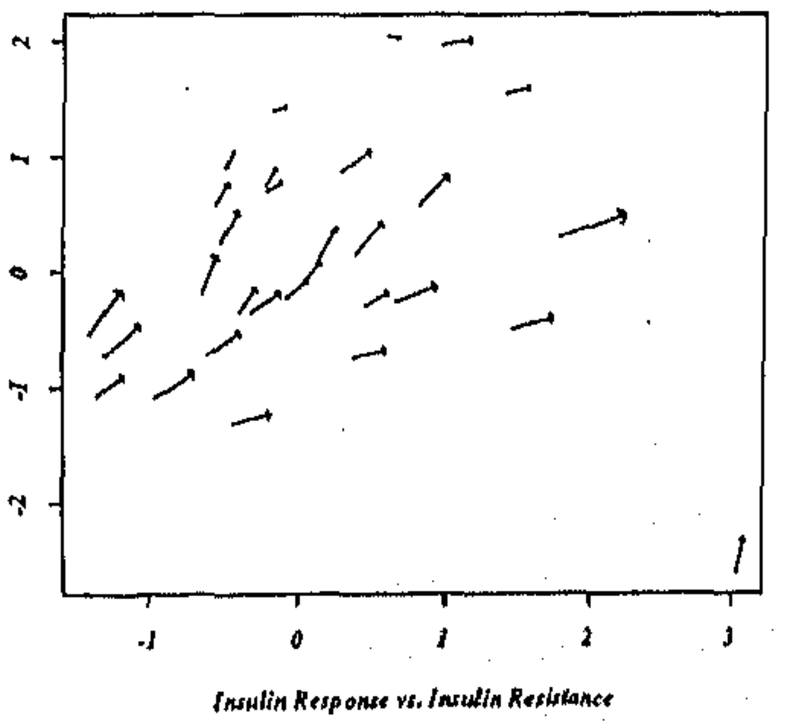
Fasting Glucose vs. Insulin Response



Glucose Resistance vs. Insulin Resistance



Fasting Glucose vs. Insulin Resistance



Insulin Response vs. Insulin Resistance

Figure 5.4: Multivariate Q-Q plot comparing normal patients with overt diabetic patients

5.4.4 L-Estimates

In the univariate set-up, linear combinations of order statistics or L-estimators have played an extremely important role in the development of robust methods for the one sample location problem. Serfling (1980) gave a detailed account of various important univariate descriptive statistics (e.g. trimmed mean, inter-quartile range etc.) by formulating them as L-statistics and derived their asymptotic properties. It is possible to extend the concept of L-estimators of univariate location to a multivariate set-up using TR l_p-quantiles in a natural way. To construct L-estimators we have to form suitable weighted averages of $\hat{Q}_n^{(\alpha,p)}(\boldsymbol{u})$'s as $\boldsymbol{u}$ varies over an appropriate subset of $B_q^{(d)}$. One has to keep in mind that a $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_q$ close to zero corresponds to a central quantile and for a $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_q$ close to one corresponds to an extreme quantile.

Suppose that μ is an appropriately chosen probability measure on $B_q^{(d)}$ supported on a subset S of $B_q^{(d)}$. Then an L-estimate of multivariate location will have the form $\int_S \hat{Q}_n^{(\alpha,p)}(\boldsymbol{u})\mu(d\boldsymbol{u})$. Specifically, if we consider $J(\boldsymbol{u})$, a bounded, real-valued continuous function defined on $B_q^{(d)}$, we may define L-estimate corresponding to the function J as,

$$\hat{\boldsymbol{\theta}}_J^{(\alpha,p)} = \int_{B_q^{(d)}} J(\boldsymbol{u})\hat{Q}_n^{(\alpha,p)}(\boldsymbol{u})\, d\boldsymbol{u} \tag{5.16}$$

By considering different forms of the function $J(\boldsymbol{u})$, one can construct various interesting descriptive statistics of the multivariate data cloud. One can define analogs of trimmed mean or inter-quartile range for a multivariate set-up. In the above set-up, if we consider S to be the l_q-ball with center at the origin and radius r, where r is a constant such that $0 < r < 1$, (i.e. is $S = \{\boldsymbol{u} | \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_q \leq r\}$) and the probability measure μ is chosen to be the uniform probability measure on S, $\int_S \hat{Q}_n^{(\alpha,p)}(\boldsymbol{u})\mu(d\boldsymbol{u})$ will be a typical definition of trimmed mean by taking $J(\boldsymbol{u}) = (\lambda(S))^{-1}1_{\{S\}}(\boldsymbol{u})$, where S is the l_q-ball of radius r as defined above and $\lambda(S)$ is the Lebesgue measure of the set S. Thus the r-trimmed multivariate mean is given by

$$\hat{\boldsymbol{\theta}}_{(r)}^{(\alpha,p)} = \frac{1}{\lambda(S)} \int_S \hat{Q}_n^{(\alpha,p)}(\boldsymbol{u})\, d\boldsymbol{u}. \tag{5.17}$$

As the transformation retransformation l_p-quantiles are equivariant under arbitrary affine transformations, the L-estimators $\hat{\boldsymbol{\theta}}_J^{(\alpha,p)}$ or the trimmed multivariate mean $\hat{\boldsymbol{\theta}}_{(r)}^{(\alpha,p)}$ are also affine equivariant. Some recent attempts to construct and study various versions of trimmed mean estimate of multivariate location using different ideas can be found in Donoho and Gasko (1992), Gordaliza (1991) and Nolan (1992). Recently, Koltchinskii (1997) showed that the geometric quantile process converges asymptotically to a Gaussian process under some suitable conditions. Using that result, he proved the asymptotic normality of the L-estimates based on non-equivariant geometric quantiles. We can establish

similar results for our TR l_p-quantile processes and derive asymptotic normality of affine equivariant L-estimates based on them.

Bibliography

ADICHIE, J. N. (1967). Estimates of regression parameters based on rank tests. *The Annals of Mathematical Statistics*, **38**, 894–904.

ADICHIE, J. N. (1978). Rank tests of subhypotheses in the general linear regression. *The Annals of Statistics*, **6**, 1012–1016.

AJNE, B. (1968). A simple test for uniformity of a circular distribution. *Biometrika*, **55**, 343–354.

AMEMIYA, T. (1982). The two stage least absolute deviations estimators. *Econometrica*, **50**, 689–711.

ANDREWS, D.F. AND HERZBERG, A.M. (1985). *Data : A Collection Of Problems From Many Fields For The Student And Research Worker.* Springer-Verlag, New York.

ARCONES, M.A., CHEN, Z. AND GINÉ, E. (1994). Estimators related to U-processes with applications to multivariate medians : Asymptotic normality. *The Annals of Statistics*, **22**, 1460–1477.

ARMSTRONG, R.D. AND KUNG, D.S. (1978). AS 132: Least absolute value estimates for a simple linear regression problem. *Applied Statistics*, **27**, 363–366.

AUBUCHON, J. C. AND HETTMANSPERGER, T. P. (1989). Rank based inference for linear models : asymmetric errors. *Statistics & Probability Letters*, 8, 97–107.

BABU, G.J. (1989). Strong representations for lad estimators in linear models. *Probability Theory and Related Fields*, **83**, 547–558.

BABU, G. J. AND RAO, C. R. (1988). Joint asymptotic distribution of marginal quantile functions in samples from a multivariate population. *Journal of Multivariate Analysis*, **27**, 15–23.

BAHADUR, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, **37**, 577–580.

BAI, Z.D., CHEN, N.R., MIAO, B.Q. AND RAO, C.R. (1990). Asymptotic theory of least distances estimate in multivariate linear models. *Statistics*, **21**, 503–519.

BAI, Z.D., RAO, C.R. AND YIN, Y.Q. (1990). Least absolute deviations analysis of variance. *Sankhyā, Series A*, **52**, 166–177.

BARNETT, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistical Society, Series A*, **139**, 318–354.

BARRODALE, I. AND ROBERTS, F.D.K. (1973). An improved algorithm for discrete L_1 linear approximation. *SIAM Journal of Numerical Analysis*, 10, 839–848.

BASSETT, G. AND KOENKER, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, **73**, 618–622.

BEDALL, F.K. AND ZIMMERMANN, H. (1979). AS 143: The mediancenter. *Applied Statistics*, **28**, 325–328.

BENNET, B.M. (1962). On multivariate sign tests. *Journal of the Royal Statistical Society, Series B*, **24**, 159–161.

BICKEL, P. J. (1964). On some alternative estimates of shift in the p-variate one sample problem. *The Annals of Mathematical Statistics*, **35**, 1079–1090.

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

BLOOMFIELD, P. AND STEIGER, W.L. (1983). *Least Absolute Deviations Theory, Applications and Algorithms*. Birkhauser, Boston.

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

BROWN, B. M. (1983). Statistical use of spatial median. *Journal of the Royal Statistical Society, Series B*, **45**, 25–30.

BROWN, B. M. (1985). Spatial median. In *Encyclopedia of Statistical Science*, New York : Wiley, vol. 8, pp. 574–575.

BROWN, B. M. AND HETTMANSPERGER, T. P. (1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Series B*, **49**, 301–310.

BROWN, B.M. AND HETTMANSPERGER, T.P. (1989). An affine invariant bivariate version of the sign test. *Journal of the Royal Statistical Society, Series B*, **51**, 117–125.

BROWN, B.M., HETTMANSPERGER, T.P., NYBLOM, J. AND OJA, H. (1992). On certain bivariate sign tests and medians. *Journal of the American Statistical Association*, **87**, 127–135.

CHAKRABORTY, B. (1997a). On multivariate median regression. Technical Report No. 9/96, Division of Theoretical Statistics & Mathematics, Indian Statistical Institute, Calcutta. Tentatively accepted in *Bernoulli*.

CHAKRABORTY, B. (1997b). On affine equivariant multivariate quantiles. Technical Report No. 17/97, Division of Theoretical Statistics & Mathematics, Indian Statistical Institute, Calcutta. (Submitted for publication)

CHAKRABORTY, B. AND CHAUDHURI, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. *Proceedings of the American Mathematical Society*, **124**, 2539–2547.

CHAKRABORTY, B. AND CHAUDHURI, P. (1997). On multivariate rank regression. *L_1-Statistical Procedures and Related Topics*, ed. Y. Dodge, IMS Lecture Notes-Monograph Series, Volume 31, pp. 399–414.

CHAKRABORTY, B. AND CHAUDHURI, P. (1998a). On an adaptive transformation and retransformation estimate of multivariate location. *Journal of the Royal Statistical Society, Series B*, *60*, 145–157.

CHAKRABORTY, B. AND CHAUDHURI, P. (1998b). On affine equivariant sign and rank tests in one and two sample multivariate problems. To appear in *Multivariate, Designs, Sample Surveys* (ed. S. Ghosh).

CHAKRABORTY, B., CHAUDHURI, P. AND OJA, H. (1998). Operating transformation retransformation on spatial median and angle test. To appear in *Statistica Sinica*.

CHATTERJEE, S.K. (1966). A bivariate sign test for location. *The Annals of Mathematical Statistics*, **37**, 1771–1782.

CHAUDHURI, P. (1992a). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *The Annals of Statistics*, **20**, 897–916.

CHAUDHURI, P. (1992b). Generalized regression quantiles : forming a useful toolkit for robust linear regression. In *L_1 Statistical Analysis and Related Methods* (Ed. Y. Dodge), North Holland : Amsterdam, pp. 169–185.

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, *91*, *862–872*.

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

CHEN, Z. (1995). Robustness of the half-space median. *Journal of Statistical Planning and Inference*, **46**, 175–181.

CHEN, Z. (1996). Bounds for the breakdown point of the simplicial median. *Journal of Multivariate Analysis*

DAVIS, J. B. AND MCKEAN, J. W. (1993). Rank based methods for multivariate linear models. *Journal of the American Statistical Association*, **88**, 245–251.

DAVIES, P.L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, **15**, 1269–1292.

DONOHO, D.L. AND GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, **20**, 1803–1827.

DRAPER, D. (1988). Rank based robust analysis of linear models I : exposition and review (with discussion). *Statistical Science*, **3**, 239–271.

EDDY, W.F. (1983). Set valued ordering of bivariate data. *Stochastic Geometry, Geometric Statistics, and Stereology*, eds. R.V. Ambartsumian and W. Weil, Leipzig: Tuebner, pp. 79–90.

EDDY, W.F. (1985). Ordering of multivariate data. *Computer Science and Statistics: The Interface*, ed. L. Billard, Amsterdam: North-Holland, pp. 25–30.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia : SIAM.

EISENHART, C. (1961). Boscovitch and the combination of observations, in *Roger Joseph Boscovitch*, ed. L.L. Whyte, Fordham University Press, New York.

FERGUSON, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.

GINI, C. AND GALVANI, L. (1929). Di talune estensioni dei concetti di media ai caratteri qualitativi. *Metron*, **8**. Partial English translation in *Journal of the American Statistical Association*, **25**, 448–450.

GORDALIZA, A. (1991). On the breakdown point of multivariate location estimators based on trimming procedures. *Statistics & Probability Letters*, **11**, 387–394.

GOWER, J.C. (1974). The mediancenter. *Journal of the Royal Statistical Society, Series C*, **23**, 466–470.

HÁJEK, J. AND ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

HALDANE, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika*, **35**, 414–415.

HAYFORD, J. F. (1902). What is the center of an area, or the center of a population? *Journal of the American Statistical Association*, **8**, 47–58.

HETTMANSPERGER, T. P. AND MCKEAN, J. W. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics*, **19**, 275–284.

HETTMANSPERGER, T. P. AND MCKEAN, J. W. (1978). Statistical inference based on ranks. *Psychometrika*, **43**, 69–79.

HETTMANSPERGER, T. P. AND MCKEAN, J. W. (1983). A geometric interpretation of inferences based on ranks in the linear model. *Journal of the American Statistical Association*, **78**, 885–893.

HETTMANSPERGER, T. P., NYBLOM, J. AND OJA, H. (1992). On multivariate notions of sign and rank. In L_1 *Statistical Analysis and Related Methods* (Editor : Y. Dodge), North Holland : Amsterdam, pp. 267–278.

HETTMANSPERGER, T. P., NYBLOM, J. AND OJA, H. (1994). Affine invariant multivariate one sample sign tests. *Journal of the Royal Statistical Society, Series B*, **56**, 221–234.

HOTELLING, H. (1929). Stability in competition. *Econom. J.*, **39**, 41–57.

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

HODGES, J.L. AND LEHMANN, E.L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, **34**, 598–611.

ISOGAI, T. (1985). Some extension of Haldane's multivariate median and its application. *Annals of the Institute of Statistical Mathematics*. **37**, 289–301.

JAN, S.L. AND RANDLES, R.H. (1994). A multivariate signed sum test for the one sample location problem. *Journal of Nonparametric Statistics*, **4**, 49–63.

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *The Annals of Mathematical Statistics*, **43**, 1449–1458.

JURECKOVA, J. (1971). Nonparametric estimation of regression coefficients. *The Annals of Mathematical Statistics*, **42**, 1328–1338.

JURECKOVA, J. (1973). Central limit theorem for Wilcoxon rank statistics process. *The Annals of Statistics*, **1**, 1046–1060.

KEMPERMAN, J. H. B. (1987). The median of a finite measure on a Banach space. *Statistical Data Analysis Based on L_1 norm and Related Methods*, ed. Y. Dodge, Amsterdam: North-Holland, pp. 217–230.

KOENKER, R. AND BASSETT, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

KOENKER, R. AND D'OREY, V. (1987). AS 229: Computing regression quantiles. *Applied Statistics*, **36**, 383–393.

KOENKER, R. AND PORTNOY, S. (1987). L estimation for linear models. *Journal of the American Statistical Association*, **82**, 851–857.

KOENKER, R. AND PORTNOY, S. (1990). M estimation of multivariate regressions. *Journal of the American Statistical Association*, **85**, 1060–1068.

KOLTCHINSKII, V. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics*, **25**, 435–477.

KOUL, H. L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *The Annals of Mathematical Statistics*, **40**, 1950–1979.

KOUL, H. L. (1970). A class of ADF tests for subhypotheses in multiple linear regression. *The Annals of Mathematical Statistics*, **41**, 1273–1281.

LEHMANN, E. L. (1963a). Asymptotically nonparametric inference : an alternative approach to linear models. *The Annals of Mathematical Statistics*, **34**, 1494–1506.

LEHMANN, E. L. (1963b). Robust estimation in analysis of variance. *The Annals of Mathematical Statistics*, **34**, 957–966.

LEHMANN, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *The Annals of Mathematical Statistics*, **35**, 726–734.

LIU, R. Y. (1988). On a notion of simplicial depth. *Proc. Nat. Acad. Sci. USA*, **85**, 1732–1734.

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, **18**, 405–414.

LIU, R.Y. AND SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, **88**, 252–259.

LOPUHAA, H. P. AND ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, **19**, 229–248.

MARDEN, J. (1998). Bivariate Q-Q plots. To appear in *Statistica Sinica*.

MARDIA, K. V. (1972). *The Statistics of Directional Data*. New York: Academic Press.

MCKEAN, J.W. AND SCHRADER, R.M. (1987). Least absolute errors analysis of variance. *Statistical Data Analysis Based on the L_1-norm and Related Methods*, ed. Y. Dodge, North-Holland, pp. 297–305.

MERCHANTS, J.A., HALPRIN, G.M., HUDSON, A.R., KILBURN, K.H., MCKENZIE, W.N., JR., HURST, D.J. AND BERMAZOHN, P. (1975). Responses to cotton dust. *Archives of Environmental Health*, **30**, 222–229.

MÖTTÖNEN, J. AND OJA, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, **5**, 201–213.

MÖTTÖNEN, J., OJA, H. AND TIENARI, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics*, **25**, 542–552.

NARULA, S.C. AND WELLINGTON, J.F. (1977). AS 108: Multiple linear regression with minimum sum of absolute errors. *Applied Statistics*, **26**, 106–111.

NEIMIRO, W. (1992). Asymptotics for M-estimators defined by convex minimization. *The Annals of Statistics*, **20**, 1514–1533.

NIINIMAA, A. AND OJA, H. (1995). On the influence functions of certain bivariate medians. *Journal of the Royal Statistical Society, Series B*, **57**, 565–574.

NIINIMAA, A., OJA, H. AND NYBLOM, J. (1992). AS 277: The Oja bivariate median. *Applied Statistics*, **41**, 611–633.

NOLAN, D. (1992). Asymptotics for multivariate trimming. *Stochastic Processes and Their Applications*, **42**, 157–169.

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, **1**, 327–332.

OJA, H. AND NYBLOM, J. (1989). Bivariate sign tests. *Journal of the American Statistical Association*, **84**, 249–259.

OJA, H., NIINIMAA, A. AND TABLEMAN, M. (1990). The finite sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. *Statistics & Probability Letters*, **10**, 325–328.

PLACKETT, R. L. (1976). Comment on "The ordering of multivariate data", by V. Barnett. *Journal of the Royal Statistical Society, Series A*, **139**, 344–346.

PURI, M. L. AND SEN, P. K. (1969). A class of rank order tests for a general linear hypothesis. *The Annals of Mathematical Statistics*, **40**, 1325–1343.

PURI, M.L. AND SEN, P.K. (1971). *Nonparametric methods in multivariate analysis*. Wiley, New York.

PURI, M. L. AND SEN, P. K. (1973). A note on asymptotically distribution free tests for subhypotheses in multiple linear regression. *The Annals of Statistics*, **1**, 553–556.

RANDLES, R.H. (1989). A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, **84**, 1045–1050.

RAO, C. R. (1988). Methodology based on the L_1-norm in statistical inference. *Sankhyā, Series A*, **50**, 289–313.

REAVEN, G.M. AND MILLER, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, **16**, 17–24.

REISS, R.D. (1989). *Approximate distributions of order statistics with applications to nonparametric statistics*. New York: Springer.

ROUSSEEUW, P.J. AND LEROY, R. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

ROUSSEEUW, P. J. AND RUTS, I. (1996). AS 307: Bivariate location depth. *Applied Statistics*, **45**, 153–168.

RUPPERT, D. AND CARROLL, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **75**, 828–838.

RUTS, I. AND ROUSSEEUW, P. J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*, **23**, 153–168.

SCATES, D.E. (1933). Locating the median of the population in the united states. *Metron*, 11, 49–66.

SCHRADER, R.M. AND MCKEAN, J.W. (1987). Small sample properties of least absolute errors analysis of variance. *Statistical Data Analysis Based on the L_1-norm and Related Methods*, ed. Y. Dodge, North-Holland, pp. 307–322.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

SMALL, C. G. (1990). A survey of multidimensional medians. *International Statistical Review*, **58**, 263–277.

SRINIVASAN, K. (1995). Recent fertility trends and prospects in India. *Current Science*, **69**, 577–586.

TUKEY, J. W. (1975). Mathematics and picturing data. In *Proceedings of the International Congress of Mathematicians, Vancouver 1974* (Ed. R.D. James), vol. 2, pp. 523–531.

WEBER, A. (1909). Uber den Standort der Industrien, Tubingen. English translation by Freidrich, C.J. (1929), *Alfred Weber's Theory of Location of Industries*, University of Chicago Press.

WELSH, A.H. (1987). The trimmed mean in the linear model (with comments). *The Annals of Statistics*, **15**, 20–45.

WESOLOWSKY, G.O. (1981). A new descent algorithm for the least absolute regression problem. *Communications in Statistics*, **B10**, 479–491.

WILKS, S.S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471–494.