Bernoulli **5**(4), 1999, 683–703

On multivariate median regression

BIMAN CHAKRABORTY

Dept. of Statistics & Appl. Prob., National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Republic of Singapore. E-mail: stabc@nus.edu.sg

An extension of the concept of least absolute deviation regression for problems with multivariate response is considered. The approach is based on a transformation and retransformation technique that chooses a data-driven coordinate system for transforming the response vectors and then retransforms the estimate of the matrix of regression parameters, which is obtained by performing coordinatewise least absolute deviations regression on the transformed response vectors. It is shown that the estimates are equivariant under non-singular linear transformations of the response vectors. An algorithm called TREMMER (Transformation Retransformation Estimates in Multivariate MEdian Regression) has been suggested, which adaptively chooses the optimal data-driven coordinate system and then computes the regression estimates. We have also indicated how resampling techniques like the bootstrap can be used to conveniently estimate the standard errors of TREMMER estimates. It is shown that the proposed estimate is more efficient than the non-equivariant coordinatewise least absolute deviations estimate, and it outperforms ordinary least-squares estimates in the case of heavy-tailed non-normal multivariate error distributions. Asymptotic normality and some other optimality properties of the estimate are also discussed. Some interesting examples are presented to motivate the need for affine equivariant estimation in multivariate median regression and to demonstrate the performance of the proposed methodology.

Keywords: affine equivariance; bootstrap; efficiency; elliptically symmetric distributions; generalized variance; least absolute deviations; multiresponse linear model; standard error estimation; transformation-retransformation estimate

1. Introduction

Consider a linear regression set-up with a k-dimensional regressor $\mathbf{x}$ and a univariate response y satisfying the linear model $y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \mathbf{e}$, where our objective is to estimate and make inferences about the k-dimensional parameter vector $\boldsymbol{\beta}$ based on a set of independent observations $(y_1, \mathbf{x}_1)$, $(y_2, \mathbf{x}_2)$, $\ldots$, $(y_n, \mathbf{x}_n)$. 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. The LAD estimation technique is known to have greater antiquity than the LS method (see, for example, 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 fitting a line by minimizing the sum of absolute deviations from the points 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

1350–7265

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 the 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 them. 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 a regression analogue of a trimmed mean. The first one was originally suggested by Koenker and Bassett (1978) and uses their concept of regression quantiles. Its asymptotic behaviour is completely analogous to that of a trimmed mean. The second method uses residuals from a preliminary estimator, and its asymptotic behaviour heavily depends on that preliminary estimate. Welsh (1987) proposed another analogue of trimmed mean using the 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 now well known that the LAD regression problem can be formulated as a linear programming problem, and, as a result, several good algorithms are available for computing 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 LAD estimators and established the strong consistency and asymptotic normality of the estimates. Subsequently McKean and Schrader (1987), Schrader and McKean (1987) and Bai *et al*. (1990b) 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 a '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 of a very general result obtained by Neimiro (1992).

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

Example 1. The Biological Sciences Division of the Indian Statistical Institute, Calcutta, collected data on blood pressures of 40 Marwari females residing in the Burrabazar area of Calcutta (see Table 1). It is well known to physiologists that arterial pressure tends to increase

Table 1. Systolic and diastolic blood pressure and age of Marwari females 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

with age. 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 ethnic status. 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 age is an important factor in deciding what should be a normal blood pressure.

We are interested here in finding an empirical relationship between systolic and diastolic blood pressures and age for a normal Marwari woman residing in Calcutta. Now in Figure 1, which shows the scatterplots of systolic and diastolic blood pressure 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 (for example, 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 between the blood pressures and 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 the multiresponse linear regression set-up and suggested the use of univariate LAD regression for each coordinate of the response vector. He has shown that under simple conditions that estimate

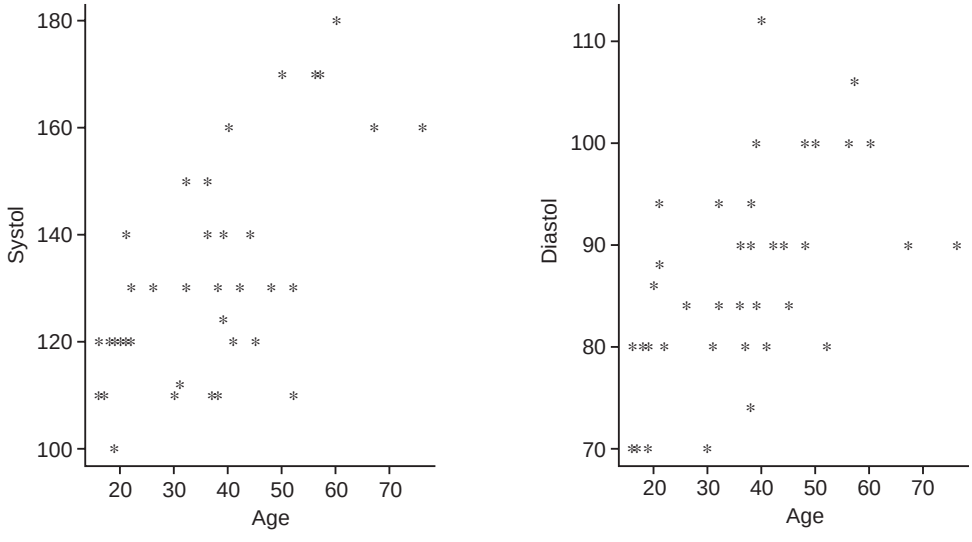

Figure 1. The systolic and diastolic blood pressures against age, for 40 women, showing high variability, with some possible outliers

is asymptotically normal, but the problem with his estimation technique is that it does not take into account the interdependence of the coordinates of the response vector, and it may not always be wise to ignore correlations that exist among different response variables. Another approach to generalizing LAD estimation in the multivariate response problem is due to Bai *et al*. (1990a), who extended the notion of spatial median (cf. Brown 1983; Chaudhuri 1992; Haldane 1948) in the regression set-up, and obtained their estimate by minimizing $\sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\|$ with respect to $\boldsymbol{\beta}$ (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 non-singular linear transformations of the response vector. In another generalization, Koenker and Portnoy (1990) suggested M-estimation in the multiresponse 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 detail. It is worth noting here (see also Chakraborty and Chaudhuri 1998) that Bickel (1964) raised an important issue related to the efficiency of the vector of coordinatewise medians. He pointed out that the performance of the vector of medians becomes really poor in the presence of high correlations among the components of the data vector. He concluded that the reason behind this abysmal performance of the vector of medians may be a lack of affine equivariance. Similar feelings have been expressed by other authors (see Brown and Hettmansperger 1987; 1989), who suggested some affine equivariant multivariate location estimates.

Chakraborty and Chaudhuri (1998) investigated in detail connections between the affine

equivariance (or lack of it) and asymptotic efficiency (or loss of it) of multivariate location estimates when the components of the data vector are correlated. Let us consider the blood pressures data discussed in Example 1. It is fairly obvious from Figure 2 that there exists high correlation between the systolic and diastolic pressures of a person as one would expect, and instead of using the LAD regression for each of the two pressure measurements separately, we need to use some affine equivariant approach.

Example 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 16 most populated states of India.

Our interest here is in exploring the relationships between TFR and IMR on the one hand, and FLR on the other, over different years and regions. As the data set contains only 48 observations, we amalgamated the states into four regions – 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 3). So in this situation too the asymptotic efficiency of the non-equivariant coordinatewise 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 interdependence between TFR and IMR.

In this paper, we propose and investigate in detail a technique for estimating parameters in linear models with multivariate response, which will lead to estimates that are equivariant under non-singular linear transformations of the response vector. First observe that the LAD

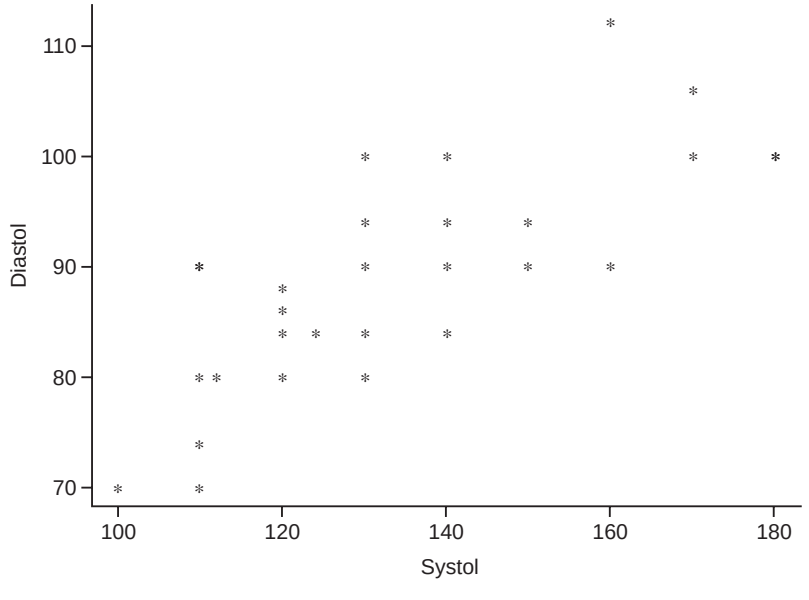

Figure 2. Systolic blood pressure against diastolic blood pressure

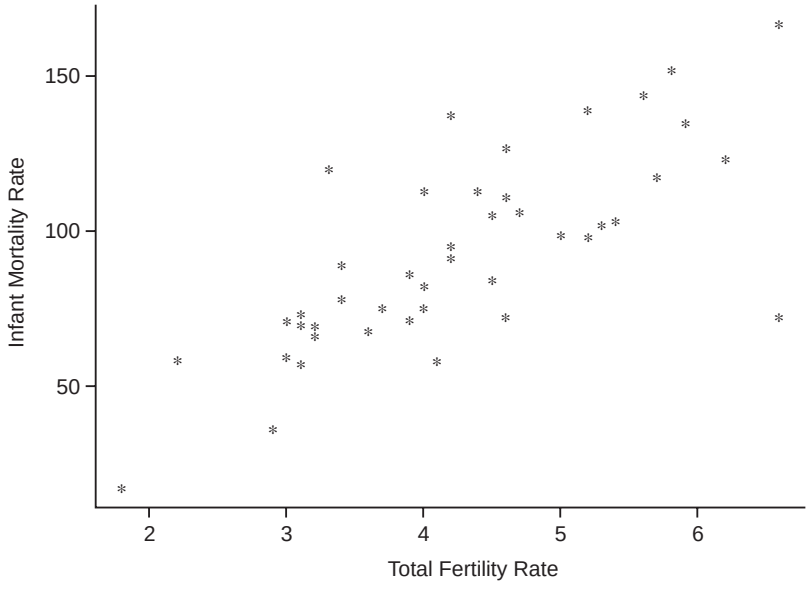

Figure 3. Total fertility rate against infant mortality rate, showing high correlation

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 the 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 analogue of the median in several dimensions. Now, there are several definitions of the multivariate median available in the literature (for detailed reviews, see Chaudhuri 1992, Small 1990). The vector of coordinatewise medians lacks the property of equivariance even under orthogonal transformations, and its regression analogue, which is the coordinatewise LAD estimates, has the same drawback. The 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*. (1990a). 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 analogue, which will be equivariant under non-singular linear transformations of the response. However, none of these regression analogues 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. Chakraborty and Chaudhuri (1996) proposed a version of the multivariate median following the idea of 'data-driven coordinate systems' introduced by Chaudhuri and Sengupta (1993), which is computationally much less intensive, and we consider here the regression analogue of their transformation-retransformation median. We will demonstrate that the proposed

estimate outperforms the matrix of coordinatewise 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 the finite-sample variance-covariance matrix of the estimates.

In Section 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. Then we demonstrate, with two real examples, the performance of the procedure. In Section 3, we establish some important asymptotic results about the proposed estimate and demonstrate some optimal properties of the adaptive transformation-retransformation median regression estimates. All proofs are postponed to the Appendix.

2. Description and computation of estimates

Consider the following multiresponse linear model:

$$\mathbf{y}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{e}_i, \qquad i = 1, \ldots, n$$

where the $\mathbf{y}_i$ are $d \times 1$ response vectors, the $\mathbf{x}_i$ are $k \times 1$ dimensional vectors of explanatory variables, the $\mathbf{e}_i$ are d-dimensional error vectors, and $\boldsymbol{\beta}$ is a $k \times d$ matrix of parameters. We assume that the $\mathbf{e}_i$ 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 non-singular $d \times d$ matrix $\mathbf{A}$, the transformation that maps $\mathbf{y}_i$ into $\mathbf{A}\mathbf{y}_i$, for $1 \leqslant i \leqslant n$, 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 data-based transformation-retransformation is to form an appropriate 'data-driven coordinate system' (see also Chakraborty *et al*. 1998) 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 the transformed response vectors. Finally, the estimates of regression parameters are retransformed so as to express everything in terms of original coordinate system (see also Chakraborty and Chaudhuri 1996; 1998). 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 obtain a valid coordinate system, these d points must satisfy some non-singularity condition.

Let us define

$$A_n = \{\mathbf{a}\colon \mathbf{a} \subset \{1, 2, \ldots, n\} \text{ and } \#\{i\colon i \in \mathbf{a}\} = k\},$$
$$B_n = \{\mathbf{b}\colon \mathbf{b} \subset \{1, 2, \ldots, n\} \text{ and } \#\{i\colon i \in \mathbf{b}\} = d\},$$
$$S_n = \{\alpha = \mathbf{a} \cup \mathbf{b}\colon \mathbf{a} \in A_n \quad \mathbf{b} \in B_n, \quad \mathbf{a} \cap \mathbf{b} = \phi\}.$$

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 $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$, and $\mathbf{Z}(\alpha)$ be the $d \times k$ matrix whose columns are the vectors $\mathbf{y}_{i_1}, \ldots \mathbf{y}_{i_k}$. 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 $\mathbf{y}_{j_1} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{x}_{j_1}, \ldots, \mathbf{y}_{j_d} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{x}_{j_d}$. If the error vectors $\mathbf{e}_i$ happen to be i.i.d. random vectors with a common probability distribution, which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, the matrix $\mathbf{E}(\alpha)$ will be invertible with probability one. Then define transformed response vectors $\mathbf{z}_j^{(\alpha)} = \{\mathbf{E}(\alpha)\}^{-1}\mathbf{y}_j$ for $1 \leqslant j \leqslant n$ with $j \notin \alpha$. Let $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ be the matrix of parameter estimates obtained by regressing each coordinate of the d-dimensional transformed vectors $\mathbf{z}_i^{(\alpha)}$ separately on the $\mathbf{x}_i$ for $1 \leqslant i \leqslant n$ and $i \notin \alpha$ using the LAD method. Then define the transformation-retransformation estimate of the parameter matrix as $\hat{\boldsymbol{\beta}}_n^{(\alpha)} = \hat{\mathbf{\Gamma}}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^{\mathrm{T}}$. Note that $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ is obtained by retransforming the earlier $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ by the linear transformation $\mathbf{E}(\alpha)$. The following theorem asserts the equivariance of $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ under non-singular transformations of the response vector and the regression equivariance of it.

Theorem 2.1 *For a fixed $\alpha \in S_n$, let $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ be the estimated matrix of parameters based on the data points $(\mathbf{y}_1, \mathbf{x}_1), (\mathbf{y}_2, \mathbf{x}_2), \ldots, (\mathbf{y}_n, \mathbf{x}_n)$ as described above.*

(i) Suppose that $\mathbf{A}$ is a fixed $d \times d$ non-singular matrix. Then the transformation-retransformation estimate computed from $(\mathbf{A}\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{A}\mathbf{y}_n, \mathbf{x}_n)$ in the same way as above (using the same index set α) will be $\hat{\boldsymbol{\beta}}_n^{(\alpha)}\mathbf{A}^{\mathrm{T}}$.

(ii) Suppose that the response vectors $\mathbf{y}_i$ are transformed to $\mathbf{y}_i - \mathbf{G}^{\mathrm{T}}\mathbf{x}_i$ for $i = 1, \ldots, n$, where $\mathbf{G}$ is a fixed $k \times d$ matrix. Then the transformation-retransformation estimate will be transformed to $\hat{\boldsymbol{\beta}}_n^{(\alpha)} - \mathbf{G}$.

From the definition of the transformation-retransformation estimate of the matrix of parameters and from Theorem 2.1, we make the following simple observations:

Observation 1. If $k = 1$ and the regressors $x_i = 1$, for $1 \leqslant i \leqslant n$, then our problem reduces to estimation of the multivariate median of the observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, and our estimation procedure leads to the transformation-retransformation multivariate median introduced in Chakraborty and Chaudhuri (1996).

Observation 2. The transformation-retransformation estimate of the parameter matrix is obtained as the minimizer of $\sum_{i \notin \alpha} |\{\mathbf{E}(\alpha)\}^{-1}(\mathbf{y}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)|$ with respect to $\boldsymbol{\beta} \in \mathbb{R}^{k \times d}$, where $|\cdot|$ denotes the l_1-norm in $\mathbb{R}^d$. This implies that the estimate $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ is equivariant under linear reparametrization of the design points $\mathbf{x}_i$. In other words, if we transform our regressor vectors $\mathbf{x}_i$ to $\mathbf{B}\mathbf{x}_i$, for $1 \leqslant i \leqslant n$, where $\mathbf{B}$ is a $k \times k$ nonsingular matrix, then our estimate is transformed to $(\mathbf{B}^{\mathrm{T}})^{-1}\hat{\boldsymbol{\beta}}_n^{(\alpha)}$.

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

$$\mathbf{y}_i = \boldsymbol{\gamma} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{e}_i, \qquad i = 1, \ldots, n.$$

If the $\mathbf{y}_i$ are transformed to $\mathbf{y}_i + \mathbf{b}$, where $\mathbf{b}$ is a d-dimensional vector, then by (ii) of

Theorem 2.1 our transformation-retransformation estimate $\hat{\boldsymbol{\gamma}}_n^{(\alpha)}$ will be transformed to $\hat{\boldsymbol{\gamma}}_n^{(\alpha)} + \mathbf{b}$ and $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ will remain unchanged.

There are several linear models with multivariate responses discussed in the existing literature, of which seemingly unrelated regression (SUR) models (Zellner 1962) deal with different sets of explanatory variables for different responses and are immensely popular among the econometricians. But to keep things simple, we have considered the multivariate regression model with the same explanatory variables for all responses. This simple model also has a lot of applications, some of which are demonstrated by our Examples 1 and 2. Multivariate analysis of variance problems can also be formulated in terms of this model. It is interesting to note that this transformation-retransformation strategy is a general tool for constructing affine equivariant estimates out of non-equivariant estimates whatever the regression model may be. In particular, it is also possible to suitably modify our transformation-retransformation strategy for SUR models.

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 $\boldsymbol{\beta}$. 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 $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$, 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 $\mathbf{e}$ is elliptically symmetric with a density of the form $\{\det(\boldsymbol{\Sigma})\}^{-1/2} f(\mathbf{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e})$ where $\boldsymbol{\Sigma}$ is a $d \times d$ positive definite matrix, and $f(\mathbf{e}^{\mathrm{T}}\mathbf{e})$ is a spherically symmetric density on $\mathbb{R}^d$, we have a nice simple form for the asymptotic generalized variance of the estimate $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ for a given α as given in the following theorem. Let us write $\{\boldsymbol{\Sigma}^{-1/2}\mathbf{E}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$, 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 $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$.

Theorem 2.2. *Fix $\alpha \in S_n$. Assume that the density f is such that any univariate marginal g of the spherically symmetric density $f(\mathbf{e}^{\mathrm{T}}\mathbf{e})$ is differentiable and positive at zero, and $\max_{i \notin \alpha} \mathbf{x}_i^{\mathrm{T}} \{\sum_{j \notin \alpha} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}\}^{-1} \mathbf{x}_i$ converges to zero as n tends to infinity. Then, as n tends to infinity, the conditional distribution of $\{\sum_{j \notin \alpha} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}\}^{1/2} (\hat{\boldsymbol{\beta}}_n^{(\alpha)} - \boldsymbol{\beta})$, given the $\mathbf{e}_i$ with $i \in \alpha$,* converges weakly to a multivariate normal distribution with zero mean and $c\boldsymbol{\Sigma}^{1/2}\mathbf{V}(\alpha)\boldsymbol{\Sigma}^{1/2} \otimes \mathbf{I}_k$ *as the variance matrix. Here $c = \{2g(0)\}^{-2}$, $\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^{\mathrm{T}}\}^{-1}$,* 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 ith and the jth row of $\mathbf{J}(\alpha)$. We denote by $\otimes$ the usual Kronecker product, and $\mathbf{I}_k$ is the identity matrix of dimension $k \times k$.*

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

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

Corollary 2.3. *Suppose that the conditions on the distribution of the error vector* $\mathbf{e}$ *stated in Theorem 2.2 are satisfied, and assume that* $n^{-1}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ *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* $\mathbf{e}_i$ *with* $i\in\alpha$, *converges weakly to a multivariate normal distribution with zero mean and* $c\mathbf{\Sigma}^{1/2}\mathbf{V}(\alpha)\mathbf{\Sigma}^{1/2}\otimes\mathbf{Q}^{-1}$ *as the variance matrix, where c and* $\mathbf{V}(\alpha)$ *are as in Theorem 2.2.*

Under the assumptions of Corollary 2.3, the expression of the asymptotic generalized variance becomes

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

The following fact, which directly follows from Theorem 3.2 of Chakraborty and Chaudhuri (1996), 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 2.4. *For the matrices* $\mathbf{D}(\alpha)$ *and* $\mathbf{J}(\alpha)$ *defined above, we have* $\det\{\mathbf{D}(\alpha)\}\geqslant[\det\{\mathbf{J}(\alpha)\}]^2$, *so that* $\det\{\mathbf{V}(\alpha)\}\geqslant 1$. *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 $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$. The above-mentioned expression for generalized variance involves the scale matrix $\mathbf{\Sigma}$, which is in general unknown. We will need a consistent affine equivariant estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$, and then we can transform the $\mathbf{y}_i$ to $\hat{\mathbf{\Sigma}}^{-1/2}\mathbf{y}_i$, for $1\leqslant i\leqslant n$, 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.

2.2. TREMMER algorithm

1. Obtain a consistent and affine equivariant estimate $\hat{\mathbf{\Sigma}}$ of the scale matrix $\mathbf{\Sigma}$ associated with the distribution of the random error $\mathbf{e}$ from the data $(\mathbf{y}_1,\mathbf{x}_1),\ldots,(\mathbf{y}_n,\mathbf{x}_n)$.
2. Transform all the response vectors $\mathbf{y}_i$ to $\hat{\mathbf{\Sigma}}^{-1/2}\mathbf{y}_i$ for $1\leqslant i\leqslant n$. Then fix a subset $\alpha\in S_n$ and compute $\hat{v}(\alpha)$ as given above; this appears in the expression for the asymptotic generalized variance of the estimate $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$.
3. Minimize $\hat{v}(\alpha)$ with respect to $\alpha\in S_n$. Call this $\hat{\alpha}$. One can reduce the amount of computation time required for searching for the optimal α by stopping whenever $\hat{v}(\alpha)$ is sufficiently close to 1 because we know from Fact 2.4 that the lower bound for $v(\alpha)$ is 1.

This approximation makes the algorithm very fast.

4. Form the matrix $\mathbf{W}(\hat{\alpha})$ with columns $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$, where $i_1, \ldots, \mathrm{i}_k$ are the first k elements of the subset $\hat{\alpha}$ and also form the matrix $\mathbf{Z}(\hat{\alpha})$ whose columns are $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_k}$. Then construct the transformation matrix $\mathbf{E}(\hat{\alpha})$ with columns $\mathbf{y}_{j_1} - \mathbf{Z}(\hat{\alpha})\{\mathbf{W}(\hat{\alpha})\}^{-1}\mathbf{x}_{j_1}$, $\ldots, \mathbf{y}_{j_d} - \mathbf{Z}(\hat{\alpha})\{\mathbf{W}(\hat{\alpha})\}^{-1}\mathbf{x}_{j_d}$ where $j_1, \ldots, \mathrm{j}_d$ are the last d elements of $\hat{\alpha}$.

5. Transform all response vectors $\mathbf{y}_i$ to $\mathbf{z}_i^{(\hat{\alpha})} = \{\mathbf{E}(\hat{\alpha})\}^{-1}\mathbf{y}_i$, for $i \notin \hat{\alpha}$. Compute the coordinatewise LAD estimate $\hat{\mathbf{\Gamma}}_n^{(\hat{\alpha})}$ of the matrix of parameters by regressing the $\mathbf{z}_i^{(\hat{\alpha})}$ on the $\mathbf{x}_i$, for $i \notin \hat{\alpha}$. Then retransform that matrix to obtain the TREMMER estimate as $\hat{\boldsymbol{\beta}}_n^{(\hat{\alpha})} = \hat{\mathbf{\Gamma}}_n(\hat{\alpha})\{\mathbf{E}(\hat{\alpha})\}^{\mathrm{T}}$.

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 of 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 an approach lies primarily in the fact that there does not exist an affine equivariant square root of the usual estimates of the $\mathbf{\Sigma}$ 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 the TREMMER algorithm) the matrix $[\mathbf{E}(\alpha)][\mathbf{E}(\alpha)]^{\mathrm{T}}$ provides an estimate of the scale matrix $\mathbf{\Sigma}$ up to some scalar multiple. The advantage of using $\mathbf{E}(\alpha)$ is that it leads to affine equivariance.

In step 1, we have to use a consistent and affine equivariant estimate of the scale matrix $\mathbf{\Sigma}$. As the methodology is quite general, one may use any estimate with those properties and it is up to the user to select a proper estimate for his/her problem. Depending on the nature of the problem, one may use robust estimates of $\mathbf{\Sigma}$ (see, Davies 1987), but in general constructing such robust estimates of $\mathbf{\Sigma}$ is computationally intensive, and if it is not extremely necessary, then one may use the variance-covariance matrix of ordinary LS residuals as an affine equivariant, consistent estimate of $\mathbf{\Sigma}$. The robustness of $\mathbf{\Sigma}$ along with the geometry of the data cloud plays an interesting role in the robustness of our estimate, which we propose to discuss elsewhere.

Note that once the matrix $\mathbf{E}(\hat{\alpha})$ is formed, the computation of $\hat{\boldsymbol{\beta}}_n^{(\hat{\alpha})}$ is straightforward as it requires solving 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 transformation, we have used resampling techniques such as the bootstrap (see also Chakraborty and Chaudhuri 1998). To implement the bootstrap, one chooses the transformation matrix adaptively first, and then, fixing that transformation matrix, one transforms all the $\mathbf{y}_i$ to obtain the $\mathbf{z}_i^{(\hat{\alpha})}$ as before. Then one computes $\hat{\mathbf{\Gamma}}_n^{(\hat{\alpha})}$ and retransforms it to obtain $\hat{\boldsymbol{\beta}}_n^{(\hat{\alpha})}$. The sampling variation of $\hat{\boldsymbol{\beta}}_n^{(\hat{\alpha})}$ is estimated by resampling from the pairs $(\mathbf{y}_i, \mathbf{x}_i)$, for $1 \leqslant i \leqslant n$, $i \notin \hat{\alpha}$, and calculating the TREMMER estimate of $\boldsymbol{\beta}$ 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. It takes only a few seconds to produce such an estimate of the conditional sampling variation of $\hat{\boldsymbol{\beta}}_n^{(\hat{\alpha})}$ based on 10 000 (say) bootstrap samples on a 486 PC. We next illustrate the procedure with Examples 1 and 2.

Example 1 (continued). Table 2 gives the TREMMER estimates of the regression coefficients, and the 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 non-equivariant LAD estimates of the regression parameters and estimated the generalized variances of both for comparison. To compare two multidimensional estimates, Bickel (1964) defined the measure of efficiency as the pth 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 efficiency of TREMMER estimate we have taken the fourth root of the ratio of the generalized variances of the TREMMER estimate and coordinatewise LAD estimate. The efficiency estimated from 10 000 bootstrap replications turns out to be 1.145 365. Figure 4 shows the TREMMER lines on the scatterplots of systolic and diastolic pressures against age.

Example 2 (continued). 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. IMR is defined as the number of deaths of children below age 1 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. FLR is defined as the percentage of literate females among those females aged 7 years and above. Our interest is to see the effect of FLR and time on TFR and IMR. Table 3 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 3, we see that both 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 the South tends to have the lowest TFR and IMR levels. In this example, we again observed that the TREMMER estimate is more efficient than the

Table 2. TREMMER estimates

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

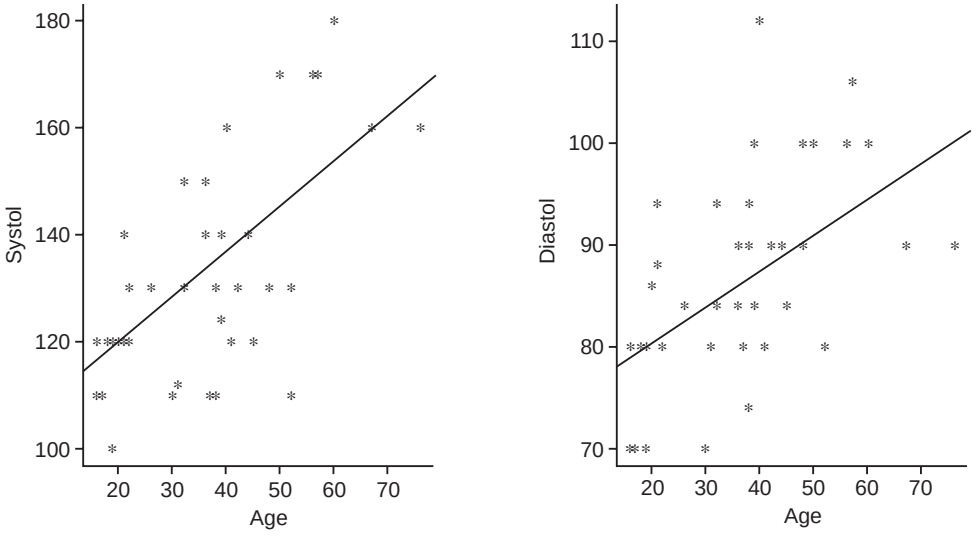

Figure 4. TREMMER regression lines superimposed on plots of blood pressure against age

Table 3. TREMMER estimates for demographic data

	Regional effects				Coeff. of Time	Coeff. of FLR
	North	East	West	South		
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)

coordinatewise LAD estimates, the efficiency being 1.456 471 as estimated from 10 000 bootstrap replications.

3. Asymptotic optimality properties of TREMMER

In this section, we will discuss the asymptotic performance of the adaptive TREMMER estimate and establish some efficiency results. For this purpose we impose some conditions on the $\mathbf{x}_i$.

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 $\|\{\mathbf{W}(\alpha)\}^{-1}\mathbf{x}_i\| \leqslant M$

for all $i \in \alpha$ and all n sufficiently large.

Condition B. The density h of a d-dimensional random vector $\mathbf{e}$ is spherically symmetric and satisfies

$$\int_{\mathbb{R}^{d\times k}} \left\{ h\left(\sum_{i=1}^{k} a_i \mathbf{e}_i \right) \right\}^d \prod_{i=1}^{k} h(\mathbf{e}_i)\, d\mathbf{e}_i < \infty$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_k$ are independent and identically distributed with common density h and the a_i 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 $\max_{1\leqslant i\leqslant n} \mathbf{x}_i^{\mathrm{T}} \{\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}\}^{-1} \mathbf{x}_i \to 0$ 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 3.1. *Suppose that the $\mathbf{x}_i$ are independent and identically distributed random variables satisfying $n^{-1}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \mathbf{Q}$ as $n \to \infty$. Then the probability of the event that Condition A holds goes to one as n tends to infinity.*

Suppose that $\alpha^* \in S_n$ minimizes $\det\{\mathbf{V}(\alpha)\}$ $(= \upsilon(\alpha)$, say) which is defined in Theorem 2.2, when the scale matrix $\mathbf{\Sigma}$ is known.

Theorem 3.2. *Assume that the $\mathbf{e}_i$ are independent and identically distributed with a common elliptically symmetric distribution $\{\det(\mathbf{\Sigma})\}^{-1/2} f(\mathbf{e}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e})$ such that the spherically symmetric density $h(\mathbf{e}) = f(\mathbf{e}^{\mathrm{T}}\mathbf{e})$ satisfies Condition B, all of its univariate marginal g is differentiable and positive at 0, $\mathbf{\Sigma}$ is a $d \times d$ positive definite matrix, and the $\mathbf{x}_i$ satisfy Condition A. Then $\upsilon(\alpha^*)$ converges to one in probability as n tends to infinity.*

Theorem 3.2 implies that if the scale matrix $\mathbf{\Sigma}$ happens to be known and the adaptive selection of $\alpha^* \in S_n$ is done using that known $\mathbf{\Sigma}$, the conditional generalized variance of the resulting adaptive TREMMER estimate tends to the lower bound established in Fact 2.4. However, in practice $\mathbf{\Sigma}$ is unknown, and we will estimate it by a consistent and affine equivariant estimate $\hat{\mathbf{\Sigma}}$ when we minimize $\hat{\upsilon}(\alpha)$ to obtain $\hat{\alpha}$. The next theorem tells us that the difference between $\upsilon(\hat{\alpha})$ and $\upsilon(\alpha^*)$ is asymptotically negligible.

Theorem 3.3. *Under the assumptions of the previous theorem, $\upsilon(\hat{\alpha}) - \upsilon(\alpha^*)$ converges in probability to zero as n tends to infinity.*

Theorems 3.2 and 3.3 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 $\mathbf{\Sigma}$ is known. If $\mathbf{\Sigma}$ is unknown, with a consistent and equivariant estimate of $\mathbf{\Sigma}$, we can choose a subset $\hat{\alpha} \in S_n$ such that $v(\hat{\alpha})$ also attains its lower bound of 1 asymptotically. Thus for n sufficiently large, we will be able to get hold of an $\hat{\alpha}$ such that $v(\hat{\alpha}) < 1 + \varepsilon$, for any $\varepsilon > 0$. Any $\alpha \in S_n$, for which $v(\alpha) < 1 + \varepsilon$, will produce an estimate with conditional asymptotic generalized variance close to $[(c/n)^d \det(\mathbf{\Sigma})]^k [\det(\mathbf{Q})]^{-d}$, where $\mathbf{Q}$ is the positive definite limit of $n^{-1}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ as n tends to infinity. From the asymptotic result obtained by Rao (1988), it can be seen that the asymptotic generalized variance of the coordinatewise LAD estimates of the parameter matrix is $[(c/n)^d \det(\mathbf{\Gamma})]^k [\det(\mathbf{Q})]^{-d}$, where the (i, j)th element of $\mathbf{\Gamma}$ 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 $\mathbf{\Sigma}$ and c is as defined earlier. Following the line of arguments used in the proof of Fact 2.4 in Chakraborty and Chaudhuri (1996), it is easy to see that $\det(\mathbf{\Gamma}) \geqslant \det(\mathbf{\Sigma})$, and equality holds only if $\mathbf{\Sigma}$ is a diagonal matrix. If the asymptotic efficiency of two competing estimates of the $k \times d$ 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 the non-equivariant coordinatewise LAD estimate is always greater than or equal to one. Further, Theorems 3.2 and 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 $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ will be more (or less) efficient than the ordinary LS estimate depending on whether the tail of the univariate marginal g of the spherically symmetric density $f(\mathbf{e}^{\mathrm{T}}\mathbf{e})$ is 'heavy' (or 'light'). It is important to note that the gain in efficiency of the affine equivariant TREMMER estimates over non-equivariant coordinatewise LAD estimates is due to the issue of affine equivariance and efficiency, a point originally made by Bickel (1964); and if the procedure, on which transformation-retransformation is applied, is affine equivariant then the resulting estimate remains same and there is no efficiency gain due to transformations. 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 non-singular 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 $\mathbf{y}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \mathbf{e}_i$ we have generated the $\mathbf{e}_i$ from bivariate normal $(f(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = (2\pi)^{-1}\exp(-(\mathbf{e}^{\mathrm{T}}\mathbf{e})/2))$ and Laplace $(f(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = (2\pi)^{-1}\exp(-\sqrt{\mathbf{e}^{\mathrm{T}}\mathbf{e}}))$ distributions, with

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

We have taken $\boldsymbol{\beta} = \mathbf{0}$, $k = 2$; the first element of $\mathbf{x}_i$ is one and the second element is generated from standard univariate normal distribution. Using these $\mathbf{e}_i$, $\mathbf{x}_i$ and $\boldsymbol{\beta}$, we have generated the observations $(\mathbf{y}_i, \mathbf{x}_i)$ 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 coordinatewise 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 $\boldsymbol{\beta}$.

From Tables 4 and 5, we see that TREMMER estimates are more efficient than coordinatewise 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 coordinatewise 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 the unconditional sampling variation of TREMMER estimates with that of the coordinatewise LAD.

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

Appendix: proofs

Proof of Theorem 2.1. (i) First observe that, in view of the way the matrix $\mathbf{Z}(\alpha)$ has been constructed, if the $\mathbf{y}_i$ are transformed to $\mathbf{A}\mathbf{y}_i$, $\mathbf{Z}(\alpha)$ 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 $\mathbf{z}_i^{(\alpha)}$ remain invariant under a non-singular linear transformation of the $\mathbf{y}_i$. Hence, the estimated matrix of regression parameters $\hat{\boldsymbol{\Gamma}}_n^{(\alpha)}$ obtained by regressing each coordinate of the $\mathbf{z}_i^{(\alpha)}$ on the $\mathbf{x}_i$ using LAD separately is invariant under that transformation. Consequently $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$, which was originally defined as $\hat{\boldsymbol{\Gamma}}_n^{(\alpha)}\{\mathbf{E}(\alpha)\}^{\mathrm{T}}$, will be transformed to $\hat{\boldsymbol{\beta}}_n^{(\alpha)}\mathbf{A}^{\mathrm{T}}$.

(ii) Observe that, if the $\mathbf{y}_i$ are transformed to $\mathbf{y}_i - \mathbf{G}^{\mathrm{T}}\mathbf{x}_i$, the matrix $\mathbf{Z}(\alpha)$ will be transformed to $\mathbf{Z}(\alpha) - \mathbf{G}^{\mathrm{T}}\mathbf{W}(\alpha)$. In turn the transformation matrix $\mathbf{E}(\alpha)$ remain invariant. Also, note that the $\mathbf{z}_i^{(\alpha)}$ are transformed to $\mathbf{z}_i^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{x}_{\mathrm{i}}$. Hence, the estimated

Table 4. 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 5. 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

matrix of regression parameters $\mathbf{\Gamma}_n^{(\alpha)}$ obtained by regressing each coordinate of the $\mathbf{z}_i^{(\alpha)}$ on the $\mathbf{x}_i$ using LAD will be transformed to $\mathbf{\Gamma}_n^{(\alpha)} - \mathbf{G}\{[\mathbf{E}(\alpha)]^{\mathrm{T}}\}^{-1}$, by the regression equivariance property of LAD. Consequently $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ will be transformed to $\hat{\boldsymbol{\beta}}_n^{(\alpha)} - \mathbf{G}$. □

Proof of Theorem 2.2. In view of the equivariance of the regression estimates $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$ under non-singular linear transformations of the $\mathbf{y}_i$, it is sufficient to prove the theorem in the special case when $\mathbf{\Sigma}$ is the $d \times d$ identity matrix. Define $\mathbf{e}_i^* = \{\mathbf{E}(\alpha)\}^{-1}\mathbf{e}_i$, for $1 \leqslant i \leqslant n$ and $i \notin \alpha$, to be the transformed error vectors. Then, given the $\mathbf{e}_i$ for which $i \in \alpha$, we have as the transformed model,

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

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

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

where U_{ji} is the ith component of $\mathbf{e}_j^*$, g_i is the ith marginal density of the distribution of $\mathbf{e}_j^*$ and $\hat{\mathbf{\Gamma}}_{in}^{(\alpha)}$ and $\mathbf{\Gamma}_i$ are the ith columns of $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ and $\mathbf{\Gamma}$ respectively. Here $\mathbf{R}_n$ converges in probability to zero. By the assumption on the $\mathbf{x}_i$ stated in the theorem, the Lindeberg condition for the central limit theorem 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 $\mathbf{e}_i$ 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, the $\mathbf{e}_i^*$ with $i \notin \alpha$ are conditionally i.i.d. random vectors with common density $|\det\{\mathbf{E}(\alpha)\}|f\{\mathbf{e}^{\mathrm{T}}[\mathbf{E}(\alpha)]^{\mathrm{T}}[\mathbf{E}(\alpha)]\mathbf{e}\}$. Let $r_1, \ldots, r_d$ be the diagonal entries of $\mathbf{R}(\alpha)$. In view of the representation discussed above, the conditional distribution of $\{\sum_{j\notin\alpha}\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}\}^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$ will converge weakly to a kd-variate normal distribution with zero mean, and limiting variance matrix $\mathbf{S}(\alpha) \otimes \mathbf{I}_k$. Here the matrix $\mathbf{S}(\alpha)$ is such that its ith diagonal entry is cr_i^2, and, for $i \neq j$, its (i, j)th element is $4cr_ir_j\{\Pr(U_{lj} < 0 \text{ and } U_{li} < 0) - \frac{1}{4}\}$. U_{li} and U_{lj} are the ith and the jth components of $\mathbf{e}_l^*$ respectively. Note that we are using the fact that for a d-dimensional random vector $\mathbf{z}$ with a spherically symmetric distribution, the distribution of the random variable $\mathbf{a}^{\mathrm{T}}\mathbf{z}$ is the same for any $\mathbf{a} \in \mathbb{R}^d$ such that $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$. Also, since the conditional distribution of $\mathbf{e}_i^*$ is elliptically symmetric around the origin in $\mathbb{R}^d$, $\Pr\{U_{lj} < 0$ 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:

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

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

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

By straightforward algebra, the proof of the theorem is now complete. □

Proof of Corollary 2.3. The proof of this corollary follows from observing the fact that $n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ converges to a positive definite matrix $\mathbf{Q}$ which implies that $\max_{1\leqslant i\leqslant n}\mathbf{x}_i^{\mathrm{T}}\{\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}\}^{-1}\mathbf{x}_i$ converges to zero as n tends to infinity. □

Proof of Proposition 3.1. As the $\mathbf{x}_i$ are independent and identically distributed, there exists $M>0$ such that, for any $\alpha\in S_n$,

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

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

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

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

Proof of Theorem 3.2. First observe that in view of the affine equivariance of $\hat{\boldsymbol{\beta}}_n^{(\alpha)}$, it is enough to consider the case when $\boldsymbol{\Sigma}=\mathbf{I}_d$. 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 $\|\{\mathbf{W}(\alpha_i)\}^{-1}\mathbf{x}_{j_l}\|$ 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 $\mathbf{e}_{j_1}-\sum_{l=1}^{k}(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}_{j_1})\mathbf{e}_{i_l}$, where $\mathbf{w}_l^{\mathrm{T}}$ is the lth row of $\{\mathbf{W}(\alpha)\}^{-1}$. As the $\mathbf{e}_i$ are i.i.d. with spherically symmetric density h, the joint pdf of $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}, \mathbf{e}_{j_1}, \ldots, \mathbf{e}_{j_d}$ can be written as $\prod_{i\in\alpha}h(\mathbf{e}_i)$. Consider the following transformation of variables

$$\mathbf{u}_1=\mathbf{e}_{j_1}-\sum_{l=1}^{k}(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}_{j_1})\mathbf{e}_{i_l},\ \ldots,\ \mathbf{u}_d=\mathbf{e}_{j_d}-\sum_{l=1}^{k}(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}_{j_d})\mathbf{e}_{i_l}$$

$$\mathbf{u}_{d+1}=\mathbf{e}_{i_1},\ \ldots,\ \mathbf{u}_{d+k}=\mathbf{e}_{i_k}.$$

Then the joint density of $\mathbf{u}_1, \ldots, \mathbf{u}_{d+k}$ is given by

$$\prod_{i=1}^{d}h\left\{\mathbf{u}_i+\sum_{l=1}^{k}(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}_{j_i})\mathbf{u}_{d+l}\right\}\prod_{i=1}^{k}h(\mathbf{u}_{d+i})$$

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

$$\int_{\mathbb{R}^{d\times k}}\prod_{i=1}^{d}h\left\{\sum_{l=1}^{k}(\mathbf{w}_l^{\mathrm{T}}\mathbf{x}_{j_i})\mathbf{u}_{d+l}\right\}\prod_{i=1}^{k}h(\mathbf{u}_{d+i})\,\mathrm{d}\mathbf{u}_{d+1}\ \ldots\ \mathrm{d}\mathbf{u}_{d+k}$$

which exists and is positive by Condition B. Now, in view of Condition A and the continuity of h at $\mathbf{0}\in\mathbb{R}^d$, the joint density of $\mathbf{u}_1, \ldots, \mathbf{u}_d$ must remain bounded away from zero in a neighbourhood of $\mathbf{0}\in\mathbb{R}^{d\times d}$. Therefore, the probability that the columns of $\mathbf{E}(\alpha)$ will be

nearly 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 $\varepsilon > 0$,

$$\inf_{\mathbf{x}_i : i \in \alpha} \Pr[v(\alpha) < 1 + \varepsilon] = p_\varepsilon > 0.$$

Then

$$\begin{aligned}\Pr\{v(\alpha^*) \geqslant 1 + \varepsilon\} &= \Pr\{\forall \alpha \in S_n,\, v(\alpha) \geqslant 1 + \varepsilon\} \\ &\leqslant \Pr\{v(\alpha_1) \geqslant 1 + \varepsilon, \ldots, v(\alpha_{k_n}) \geqslant 1 + \varepsilon\} \\ &\leqslant (1 - p_\varepsilon)^{k_l} \to 0 \text{ as } n \to \infty. \qquad \square\end{aligned}$$

Proof of Theorem 3.3. As $\hat{\mathbf{\Sigma}}$ is a consistent estimate of $\mathbf{\Sigma}$, by the simple arguments used in Chakraborty and Chaudhuri (1998), 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

$$\sup_{\alpha \in S_n} |\hat{\mathbf{D}}(\alpha) - \mathbf{D}(\alpha)| \xrightarrow{p} 0, \tag{1}$$

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

and

$$\sup_{\alpha \in S_n} |\det\{\hat{\mathbf{D}}(\alpha)\} - \det\{\mathbf{D}(\alpha)\}| \xrightarrow{p} 0 \qquad \text{as } n \to \infty. \tag{3}$$

For $M' > 0$, define $K^n_{M'} = \{\alpha : \alpha \in S_n \text{ and } v(\alpha) \leqslant M'\}$. Then by (1), (2) and (3) 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 3.2, we have that 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 $|\hat{v}(\hat{\alpha}) - v(\hat{\alpha})| < \varepsilon$, and $|\hat{v}(\alpha^*) - v(\alpha^*)| < \varepsilon$ will imply that $|\hat{v}(\hat{\alpha}) - v(\alpha^*)| < \varepsilon$. Hence, it follows that $\hat{v}(\hat{\alpha}) - v(\alpha^*)$ converges in probability to zero, which completes the proof with previous observations. $\square$

Acknowledgements

The research presented here is partially supported by a grant from Indian Statistical Institute. The author is grateful to Probal Chaudhuri for helpful suggestions and comments. He also thanks P. Bharati for making available the data on systolic and diastolic blood pressures analysed in the paper. Finally, thanks are due to two anonymous referees for a very careful reading of an earlier version of the manuscript and raising several interesting points.

References

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

Armstrong, R.D. and Kung, D.S. (1978) AS132: Least absolute value estimates for a simple linear regression problem. *Appl. Statist*., **27**, 363–366.

Babu, G.J. (1989) Strong representations for LAD estimators in linear models. *Probab. Theory Related Fields*, **83**, 547–558.

Bai, Z.D., Chen, N.R., Miao, B.Q. and Rao, C.R. (1990a) Asymptotic theory of least distances estimate in multivariate linear models. *Statistics*, **21**, 503–519.

Bai, Z.D., Rao, C.R. and Yin, Y.Q. (1990b) Least absolute deviations analysis of variance. *Sankhyā Ser. A*, **52**, 166–177.

Barrodale, I. and Roberts, F.D.K. (1973) An improved algorithm for discrete L_1 linear approximation. *SIAM J. Numer. Anal.*, **10**, 839–848.

Bassett, G. and Koenker, R. (1978) Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc*., **73**, 618–622.

Bickel, P.J. (1964) On some alternative estimates for shift in the *p*-variate one sample problem. *The Ann. Math. Statist*., **35**, 1079–1090.

Bloomfield, P. and Steiger, W.L. (1983) *Least Absolute Deviations Theory, Applications and Algorithms*. Boston: Birkhäuser.

Brown, B.M. (1983) Statistical use of the spatial median. *J. Roy. Statist. Soc. Ser. B*, **45**, 25–30.

Brown, B.M. and Hettmansperger, T.P. (1987) Affine invariant rank methods in the bivariate location model. *J. Roy. Statist. Soc. Ser. B*, **49**, 301–310.

Brown, B.M. and Hettmansperger, T.P. (1989) An affine invariant bivariate version of the sign test. *J. Roy. Statist. Soc. Ser. B*, **51**, 117–125.

Chakraborty, B. and Chaudhuri, P. (1996) On a transformation and re-transformation technique for constructing affine equivariant multivariate median. *Proc. Amer. Math. Soc.*, **124**, 2539–2547.

Chakraborty, B. and Chaudhuri, P. (1998) On an adaptive transformation retransformation estimate of multivariate location. *J. Roy. Statist. Soc. Ser. B*, **60**, 145–157.

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

Chaudhuri, P. (1992) Multivariate location estimation using extension of *R*-estimates through *U*-statistics type approach. *Ann. Statist*., **20**, 897–916.

Chaudhuri, P. and Sengupta, D. (1993) Sign tests in multidimension: inference based on the geometry of the data cloud, *J. Amer. Statist. Assoc.*, **88**, 1363–1370.

Davies, P.L. (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, **15**, 1269–1292.

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

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

Koenker, R. and Bassett, G. (1978) Regression quantiles, *Econometrica*, **46**, 33–50.

Koenker, R. and d'Orey, V. (1987) AS229: Computing regression quantiles. *Appl. Statist*., **36**, 383–393.

Koenker, R. and Portnoy, S. (1990) M-estimation of multivariate regressions. *J. Amer. Statist. Assoc.*, **85**, 1060–1068.

Liu, R.Y. (1990) On a notion of data depth based on random simplices. *Ann. Statist*., **18**, 405–414.

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

Narula, S.C. and Wellington, J.F. (1977) AS 108: Multiple linear regression with minimum sum of absolute errors. *Appl. Statist*., **26**, 106–111.

Neimiro, W. (1992) Asymptotics for *M*-estimators defined by convex minimization. *Ann. Statist*., **20**, 1514–1533.

Oja, H. (1983) Descriptive statistics for multivariate distributions, *Statist. Probab. Lett.*, **1**, 327–332.

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

Ruppert, D. and Carroll, R.J. (1980) Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.*, **75**, 828–838.

Schrader, R.M. and McKean, J.W. (1987) Small sample properties of least absolute errors analysis of variance. In Y. Dodge (ed.), *Statistical Data Analysis Based on the* L_1*-norm and Related Methods*, pp. 307–321. Amsterdam: North-Holland.

Small, C.G. (1990) A survey of multidimensional medians. *Internat. Statist. Rev.*, **58**, 263–277.

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

Tukey, J.W. (1975) Mathematics and picturing of data. In R.D. James (ed.), *Proceedings of the International Congress of Mathematicians, Vancouver 1974*, Canadian Mathematical Congress, Montreal, Que., Vol. 2, pp. 523–531.

Welsh, A.H. (1987) The trimmed mean in the linear model (with Comments). *Ann. Statist*., **15**, 20–45.

Wesolowsky, G.O. (1981) A new descent algorithm for the least absolute regression problem. *Comm. Statist. B*, **10**, 479–491.

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

Zellner, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.*, **57**, 348–368.

Received December 1996 and revised April 1998.