L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

On multivariate rank regression

Biman Chakraborty

Indian Statistical Institute, Calcutta, India

Probal Chaudhuri

Indian Statistical Institute, Calcutta, India

Abstract: An extension of rank regression techniques to multivariate linear models is proposed and studied. Unlike the co-ordinatewise rank regression techniques considered by some earlier authors, our approach is affine equivariant, and it is based on a transformation and retransformation procedure originally developed by Chakraborty and Chaudhuri (1996, 1997) for constructing an affine equivariant version of multivariate median. Affine equivariance is expected to lead to superior statistical performance of our procedure compared to other non-equivariant procedures especially in the presence of substantial correlations among different response variables in multi-response problems. Some of the statistical properties of the proposed multivariate rank regression estimates are discussed, and a few results based on numerical investigation of the performance of these estimates are presented.

Key words: Affine equivariance, Hodges-Lehmann estimate, multivariate linear model, statistical efficiency, transformation retransformation estimates, Wilcoxon's rank scores.

AMS subject classification: Primary 62H12, 62J05; secondary 62F35.

1 Introduction: multivariate linear model and rank regression

Linear model is a widely used statistical tool for empirical analysis to understand and make inference about the nature of inter-dependence that exists among different variables in the data. Perhaps it will not be an overstatement to say that various forms of linear model pervade almost every area of applied research where statistics has its scope of being effectively used. Here our focus will be on multivariate linear models of the form

$$\mathbf{Y} = \mathbf{\Gamma X} + \mathbf{e} \ , \tag{1}$$

where $\mathbf{Y}$ is a d-dimensional column vector of dependent or response variables, $\mathbf{X}$ is a p-dimensional column vector of independent or regressor variables, and $\mathbf{\Gamma}$ is the $d \times p$ matrix of unknown coefficient parameters that determine how different regressor variables jointly influence different response variables. This matrix is to be estimated from the observed data $(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$, and the term $\mathbf{e}$ in (1) is a d-dimensional column vector of random errors representing the random deviation of a data point from the linear model. In the special case when the response is real valued (i.e. when $d = 1$), rank regression techniques have been proposed and extensively studied as an alternative 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 $\mathbf{e}$ in the linear model (1) 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 location 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.

Unfortunately most of the work documented in the existing literature on rank regression is restricted to univariate response. While many practical situations (e.g. when a medical scientist is interested in studying the relationship between the age of an individual and his/her systolic and diastolic blood pressures) do involve multi-response problems, virtually very little is available in the literature other than least squares techniques when the response $\mathbf{Y}$ in the linear model (1) happens to be multi-dimensional in nature (i.e. $d > 1$). Rao (1988) and Koenker and Portnoy (1990) considered robust estimation of parameters in multi-response linear regression problems and suggested the use of univariate least absolute deviations method for each co-ordinate of the response vector. Davis and McKean (1993) have extensively studied the co-ordinatewise 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. One serious drawback of co-ordinatewise extension of rank regression as well as that of least absolute deviations regression is that such extensions fail to take into account the inter-dependence that exists 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. Such an approach for multivariate linear models leads to parameter estimates that are not equivariant and to statistical tests that are not invariant under general affine transformation of the data, and it is known that procedures that lack affine equivariance/invariance may have very poor statistical performance in the presence of substantial correlation among the components of the response vector. This issue has been discussed and investigated by Bickel (1964), Brown and Hettmansperger (1987, 1989) and Chakraborty and Chaudhuri (1996, 1997) in the context of multivariate location problems. It will be appropriate to note here that Bai, Chen, Miao and Rao (1990) proposed to estimate $\mathbf{\Gamma}$ in the linear model (1) by minimizing the sum $\sum_{i=1}^{n} \|\mathbf{Y}_i - \mathbf{\Gamma}\mathbf{X}_i\|$ w.r.t. $\mathbf{\Gamma}$, where for a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, $\|\mathbf{x}\|$ = the usual Euclidean norm of $\mathbf{x} = (x_1^2 + x_2^2 + \ldots + x_d^2)^{1/2}$, and such an estimate of $\mathbf{\Gamma}$ can be viewed as a generalization of the notion of spatial median [see e.g. Haldane (1948), Brown (1983)] in linear models. While this leads to estimates that are equivariant under rotation or orthogonal transformation of the response vector, such estimates still lack equivariance under general affine transformation of the response. 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), Chakraborty, Chaudhuri and Oja (1997)]. 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 $\mathbf{Y}$ are measured in different scales or when the response variables have different degrees of statistical variation.

Chakraborty (1996) proposed and studied in detail an affine equivariant extension of least absolute deviations regression in multi-response linear model problems using a data driven transformation and retransformation approach, which was used earlier by Chakraborty and Chaudhuri (1996, 1997) to construct an affine equivariant version of multivariate median. Such a transformation and retransformation technique converts nonequivariant estimates into equivariant ones and thereby improves upon the performance of the estimates as measured by appropriate statistical efficiency criteria. Our goal in this paper is to use the same transformation and retransformation strategy for developing affine equivariant rank regression techniques that can be used in the analysis of data following multivariate linear models. Chakraborty (1996) gave a convenient algorithm (called TREMMER) for computing the estimate of the parameter matrix $\mathbf{\Gamma}$ in (1) and amply demonstrated how resampling strategies like the bootstrap can be invoked to estimate finite sample-variance covariance matrix of such a parameter estimate. In Section 2 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 3, 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 co-ordinatewise 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 two real data sets each with bivariate response in an attempt to demonstrate the implementation of the methodology in real data and how it outperforms some competing nonequivariant procedures. Section 4 will conclude the paper with some remarks on the issues that have become transparent in course of our present research, and there we will try to discuss briefly some of the open problems that require further research.

2 The transformation and retransformation procedure

Let us now focus our attention on the data points $(\mathbf{X}_i, \mathbf{Y}_i)$'s, which are assumed to satisfy the linear model (1). Suppose that $n > d + p$, and α is a subset of size $d+p$ of the set of indices $\{1,2,\ldots,n\}$. Following the notation used by Chakraborty (1996), we will write $\alpha = \{i_1,\ldots,i_p,j_1,\ldots,j_d\}$ and denote by $\mathbf{W}(\alpha)$ the $p \times p$ matrix whose columns are the vectors $\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_p}$ and by $\mathbf{Z}(\alpha)$ the $d \times p$ matrix whose columns are the vectors $\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_p}$. We will assume that $\mathbf{W}(\alpha)$ is invertible and form the $d \times d$ matrix $\mathbf{E}(\alpha)$ that consists of the columns $\mathbf{Y}_{j_1} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{X}_{j_1},\ldots,\mathbf{Y}_{j_d} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{X}_{j_d}$. The matrix $\mathbf{E}(\alpha)$ too is assumed to be non-singular, and we define the transformed response vectors as $\mathbf{Z}_l^{(\alpha)} = \{\mathbf{E}(\alpha)\}^{-1}\mathbf{Y}_l$ for $1 \le l \le n$ and $l \not\in \alpha$. Suppose now that we perform rank regression on each co-ordinate of $\mathbf{Z}_l^{(\alpha)}$ separately with $\mathbf{X}_l$ 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 $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$. In other words, $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ is obtained by minimizing (w.r.t. $\mathbf{\Lambda}$) a dispersion function $\mathcal{D}(\mathbf{\Lambda})$ (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 $\mathcal{D}(\mathbf{\Lambda})$ is a function of the real valued co-ordinates of the multivariate residuals $\mathbf{Z}_l^{(\alpha)} - \mathbf{\Lambda}\mathbf{X}_l$ with $1 \le l \le n$, $l \not\in \alpha$ and their ranks [see Davis and McKean (1993)]. Finally, the transformation retransformation estimate of $\mathbf{\Gamma}$ is obtained by retransforming $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ by the matrix $\mathbf{E}(\alpha)$ as follows

$$\hat{\mathbf{\Gamma}}_n^{(\alpha)} = \mathbf{E}(\alpha)\hat{\mathbf{\Lambda}}_n^{(\alpha)} \ . \tag{2}$$

In view of the definition of $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$, we now have the following result, which asserts that it is an affine equivariant estimate of $\mathbf{\Gamma}$. As a matter of fact, this result is the analog of Chakraborty's (1996) Theorem 2.1 in the context of rank regression.

Result 1 *Suppose that* $\mathbf{A}$ *is a fixed* $d \times d$ *non-singular matrix. Then the transformation retransformation estimate computed from the transformed data points* $(\mathbf{X}_1, \mathbf{A}\mathbf{Y}_1), (\mathbf{X}_2, \mathbf{A}\mathbf{Y}_2), \ldots, (\mathbf{X}_n, \mathbf{A}\mathbf{Y}_n)$ *in the same way as above (i.e. using the same index set* α*) will be* $\mathbf{A}\hat{\mathbf{\Gamma}}_n^{(\alpha)}$. *Further, if the response vector* $\mathbf{Y}_i$ *is transformed to* $\mathbf{Y}_i - \mathbf{G}\mathbf{X}_i$ *for each* $i = 1,2,\ldots,n$, *where* $\mathbf{G}$ *is a fixed* $d \times p$ *matrix, the transformation retransformation estimate gets transformed to* $\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{G}$.

Proof: In view of the construction of $\mathbf{Z}(\alpha)$, when the $\mathbf{Y}_i$'s are transformed to $\mathbf{A}\mathbf{Y}_i$'s, $\mathbf{Z}(\alpha)$ gets transformed to $\mathbf{A}\mathbf{Z}(\alpha)$, and consequently the transformation matrix $\mathbf{E}(\alpha)$ is transformed to $\mathbf{A}\mathbf{E}(\alpha)$. On the other hand, since the $\mathbf{Z}_l^{(\alpha)}$'s remain invariant under non-singular linear transformation of the $\mathbf{Y}_i$'s, so does the estimate $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$, which is obtained by performing co-ordinatewise rank regression of the $\mathbf{Z}_l^{(\alpha)}$'s on the regressor vectors $\mathbf{X}_l$'s.

Hence, the transformation retransformation estimate will be transformed to $\mathbf{A}\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ as it has been claimed in the statement of the result. Next note that when each of the $\mathbf{Y}_i$'s is transformed as $\mathbf{Y}_i - \mathbf{G}\mathbf{X}_i$, the matrix $\mathbf{Z}(\alpha)$ becomes $\mathbf{Z}(\alpha) - \mathbf{G}\mathbf{W}(\alpha)$. However, the matrix $\mathbf{E}(\alpha)$ remains unaltered after such a transformation of the response. Since each of the $\mathbf{Z}_l^{(\alpha)}$'s gets transformed as $\mathbf{Z}_l^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}\mathbf{X}_l$'s in this case, the equivariance of co-ordinatewise rank regression estimate implies that $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ will now become $\mathbf{\Lambda}_n^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}$. The proof is now complete in view of the way $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ has been formed by retransforming $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ into $\mathbf{E}(\alpha)\hat{\mathbf{\Lambda}}_n^{(\alpha)}$. □

2.1 Selection of the optimal data driven transformation

Since the estimate $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ 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. This question has been dealt with in other situations by Chakraborty and Chaudhuri (1996, 1997), Chakraborty (1996) and Chakraborty, Chaudhuri and Oja (1997), who used transformation and retransformation techniques in different multivariate estimation problems. Depending on the nature of the problem, these authors have 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 $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ is a $n^{1/2}$-consistent and asymptotically normal estimate of the parameter matrix $\mathbf{\Gamma}$ in the linear model (1).

Result 2 *Fix an α. Suppose that the distribution of the $(\mathbf{X}_i, \mathbf{Y}_i)$'s and the nature of the dispersion function $\mathcal{D}(\mathbf{\Lambda})$ are such that $n^{1/2}$-consistency and asymptotic normality of the co-ordinatewise rank regression estimates holds. For example the regularity conditions used in Davis and McKean (1993), who considered co-ordinatewise rank regression will be sufficient for this purpose. Then conditioned on α and the $(\mathbf{X}_i, \mathbf{Y}_i)$'s with $i \in \alpha$, the asymptotic distribution of $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$ 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 $(\mathbf{X}_i, \mathbf{Y}_i)$'s with $i \in \alpha$. Note that since $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ is obtained by performing co-ordinatewise rank regression of the transformed response vectors $\mathbf{Z}_l^{(\alpha)}$'s on the covariates $\mathbf{X}_l$'s, it will be a $n^{1/2}$-consistent and asymptotically normal estimate of $\{\mathbf{E}(\alpha)\}^{-1}\mathbf{\Gamma}$ under appropriate regularity conditions as assumed in the statement of the result. The proof is now complete if we recall that $\hat{\mathbf{\Gamma}}_n^{(\alpha)} = \mathbf{E}(\alpha)\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ and use the fact that linear transformation preserves multivariate normality as well as $n^{1/2}$-consistency. □

However, the conditional asymptotic dispersion matrix of $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ 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 $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$, and then select an optimal $\mathbf{E}(\alpha)$ based on this estimate. However, 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 $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ will require for different choices of α.

Suppose now that $\mathbf{e}$ has an elliptically symmetric distribution with a density of the form $\{\det(\Sigma)\}^{-1/2} f(\mathbf{e}^T\Sigma^{-1}\mathbf{e})$, 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 it was proved by Chakraborty (1996) under suitable conditions that the asymptotic generalized variance of the transformation retransformation median regression (i.e. TREMMER) estimate of $\mathbf{\Gamma}$ in the linear model (1) is minimized by choosing α to minimize the determinant of the matrix

$$\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{[\mathbf{J}(\alpha)]^T\}^{-1} \ . \tag{3}$$

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 $\mathbf{E}(\alpha)$ to achieve a similar goal 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. 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 co-ordinate system determined by the transformation matrix $\mathbf{E}(\alpha)$ such that the co-ordinate 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 (1) [see also Chakraborty and Chaudhuri (1996, 1997) and Chakraborty, Chaudhuri and Oja (1997)]. 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.

2.2 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

$$\mathcal{D}(\mathbf{\Lambda}) = \sum_{1 \le r < s \le n} \sum_{;r,s \notin \alpha} \left|(\mathbf{Z}_r^{(\alpha)} + \mathbf{Z}_s^{(\alpha)}) \; - \; \mathbf{\Lambda}(\mathbf{X}_r + \mathbf{X}_s)\right| \tag{4}$$

or

$$\mathcal{D}(\mathbf{\Lambda}) = \sum_{1 \le r < s \le n} \sum_{;r,s \notin \alpha} \left|(\mathbf{Z}_r^{(\alpha)} - \mathbf{Z}_s^{(\alpha)}) \; - \; \mathbf{\Lambda}(\mathbf{X}_r - \mathbf{X}_s)\right| , \tag{5}$$

where for a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, $|\mathbf{x}|$ = the l_1-norm of $\mathbf{x} = |x_1| + |x_2| + \ldots + |x_d|$. Note that the dispersion in (4) originates from Wilcoxon's signed rank score used in single sample location problems while that in (5) is related to the two sample Wilcoxon's rank test. The second dispersion can also 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 (1). On the other hand the dispersion function in (4) 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) and (5) 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 co-ordinatewise least absolute deviations problem, and hence the computation of the transformation retransformation estimate $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ can be easily handled by some straight forward modification of the TREMMER algorithm developed by Chakraborty (1996). One only needs to replace the original data by their pairwise averages or differences (depending on whether (4) or (5) 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 2.1 when rank regression is performed using Wilcoxon's score in a multivariate linear model with the residual having multivariate normal distribution.

Result 3 *Suppose that the residuals $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{\Gamma}\mathbf{X}_i$ 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 $\mathbf{X}_i$'s have a distribution with an associated $p \times p$ expected information matrix $E(\mathbf{X}_i\mathbf{X}_i^T) = \mathbf{Q}$ that is positive definite ensuring asymptotic normality of the co-ordinatewise rank regression estimates obtained using the dispersion function (4) or (5) [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 2.1 yields a transformation retransformation estimate $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ such that the asymptotic generalized variance of $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$ tends to its minimum possible value as n tends to infinity.*

Proof: Once again let us fix α and argue conditionally give the $(\mathbf{X}_i, \mathbf{Y}_i)$'s with $i \in \alpha$. Note that when the dispersion function (5) 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 $\mathbf{X}_i$'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 $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ using the asymptotic results in Chaudhuri (1992b), and this implies that as n tends to infinity, the limiting distribution of $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$ is multivariate normal with zero mean and a variance covariance matrix that has the form

$$\Sigma^{1/2}\{\mathbf{J}(\alpha)\}^{-1}\mathbf{H}(\alpha)\{[\mathbf{J}(\alpha)]^T\}^{-1}\Sigma^{1/2} \otimes \mathbf{Q}^{-1} \ , \qquad (6)$$

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 2.1, 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 (6) 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. □

3 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 a couple of real data sets for which there are some appropriate multi-response linear models. 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 (1) with $d = p = 2$, and the first co-ordinate of $\mathbf{X}$ was taken to be the constant 1.0 while the second co-ordinate was generated from a standard normal distribution. We chose $\mathbf{\Gamma}$ 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 $\{\det(\Sigma)\}^{-1/2} f(\mathbf{e}^T\Sigma^{-1}\mathbf{e})$. These distributions are bivariate normal, bivariate Laplace [i.e. when $f(\mathbf{e}^T\mathbf{e}) = (2\pi)^{-1}\exp(\sqrt{\mathbf{e}^T\mathbf{e}})$] and bivariate t with 3 degrees of freedom. We used the dispersion function (4) for computing the transformation retransformation estimate $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ after choosing α using the selection procedure described in Section 2.1. Let $\mathcal{E}_{ols}$ and $\mathcal{E}_{lad}$ denote the efficiencies of our estimates compared with the ordinary least squares and co-ordinatewise 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 $\mathbf{\Gamma}$ 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 co-ordinatewise least absolute deviations regression does not lead to an affine equivariant estimate of $\mathbf{\Gamma}$, $\mathcal{E}_{lad}$ 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 the following table.

Table 3.1: Values of $\mathcal{E}_{lad}$ for different choices of the residual distribution and ρ.

Residual	Values of ρ				
Distribution	0.75	0.80	0.85	0.90	0.95
Bivariate Normal	158%	174%	185%	207%	244%
Bivariate Laplace	125%	128%	144%	159%	193%
Bivariate t with 3 d.f.	121%	128%	141%	158%	189%

Analysis of Blood Pressure Data: This data was collected by the Biological Sciences Division of Indian Statistical Institute, Calcutta, and it consists of systolic and diastolic blood pressures of 40 Marwari females residing at Burrabazar area of Calcutta and their ages. It is well known to physiologists that arterial pressure increases with age, and age is considered to be a factor of prime importance in deciding what should be the normal arterial pressure of an individual. As one would expect, there is ample evidence in the data [see e.g. Chakraborty (1996) who analyzed the same data using TREMMER] for the the presence of high positive correlation between systolic and diastolic pressures, and hence one can argue in favor of using an affine equivariant procedure, which is expected to be statistically more efficient for analyzing this data set than a non-equivariant procedure such as the co-ordinatewise least absolute deviations regression. We applied our affine equivariant rank regression procedure based on the dispersion function (4) to this data and obtained the following estimated linear equations : *systolic pressure* $= 100.64 + 0.8(age)$, and *diastolic pressure* $= 74.04 + 0.32(age)$. Following Chakraborty (1996), 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 co-ordinatewise least absolute deviations regression. The coefficients of age in both the equations here are slightly larger than those obtained by Chakraborty (1996) 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 [cf. Chakraborty (1996)]. It will be appropriate to note here that for using TREMMER, Chakraborty

(1996) reported 14.5% gain in statistical efficiency over non-equivariant co-ordinatewise least absolute deviations.

Analysis of Demographic Data: This second data set consists of 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 in India. The data is given in a nicely compiled form in Srinivasan (1995). TFR is defined as the number of children born to a woman in her entire reproductive span assuming that she experiences the level of age-specific fertility rates in a given period of time, and IMR is the number of deaths of infants (i.e. children below age of one year) per thousand live births during a given period. Socio-demographic studies have strongly revealed education of women as a major determinant of visible decline in infant mortality and total fertility levels of the population. Our main interest here is in exploring the nature of dependence of TFR and IMR on FLR as well as the changes in TFR and IMR over time and their regional variations, and for this we have used a multivariate analysis of covariance type model with four regional effect parameters (corresponding to northern, southern, eastern and western regions of the country) and two covariates, namely FLR and time. Once again in view of strong correlation between TFR and IMR [see Chakraborty (1996)], any non-equivariant estimation procedure is expected to perform poorly in this case. Since here one is interested in the differences between regional effects, the dispersion function in (5) is quite appropriate. When we compared our affine equivariant procedure with non-equivariant co-ordinatewise 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, our analysis did not reveal any statistically significant regional difference in fertility and mortality rates. These findings are in conformity with the results reported in Chakraborty (1996) who analyzed the same data using TREMMER.

4 Concluding remarks and discussion

An important issue that emerges at this point is that the problem of multivariate rank regression is intrinsically related to the problem of multivariate quantiles and ranks. Readers are referred to Chaudhuri (1996) and Mottonen and Oja (1995) for a detailed review of various notions of multivariate quantiles and ranks. A particularly interesting alternative to our present approach of multivariate rank regression will be to use the ranks associated with spatial or geometric quantiles [see Chaudhuri (1996), Mottonen and Oja (1995)]. Affine equivariance can still be achieved through data driven transformation and retransformation as has been done in Chakraborty, Chaudhuri and Oja (1997), where an equivariant modification of spatial median and an invariant modification of angle test were proposed and studied based on the idea of transformation and retransformation. However, such a geometric concept of quantiles leads to vector valued ranks that are very different in nature from co-ordinatewise ranks, and one needs to redefine the multivariate analog of Jaeckel's dispersion function appropriately using some suitable notion of score functions defined for vector valued ranks.

So far we are able to prove asymptotic optimality of our procedure for selecting the subset of indices α and the associated transformation matrix $\mathbf{E}(\alpha)$ only in a very special case, i.e. for dispersions associated with Wilcoxon's rank scores, and when the residual in the linear model (1) is normally distributed. In the case of multivariate median (or least absolute deviations) regression, Chakraborty (1996) was able to show that the proposed data based selection rule for choosing the transformation matrix leads to an asymptotically optimal solution whenever the residual distribution is elliptically symmetric. The nice geometric interpretation of this selection procedure described in Section 2.1 makes us believe that its asymptotic optimality holds under much weaker and more general conditions than what we have assumed in Result 3.

Rank regression in linear models with univariate response generated fascinating research problems and innovative statistical tools for nearly three decades [see Draper (1988)]. This enriched our theoretical understanding of linear model analysis and enabled us to invent new methodology for exploring relationships present among different variables in the data. Multivariate rank regression is likely to lead us to a more fertile ground for methodological and theoretical research. As multi-response problems do arise often in practice, there seems to be a real need for a serious and extensive research of rank regression to deal with such problems.

Acknowledgements

Research presented here was partially supported by a grant from Indian

Statistical Institute.

References

[1] Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Ann. Math. Statist.* **38**, 894–904.

[2] Adichie, J. N. (1978). Rank tests of subhypotheses in the general linear regression. *Ann. Statist.* **6**, 1012–1016.

[3] Aubuchon, J. C. and Hettmansperger, T. P. (1989). Rank based inference for linear models : asymmetric errors. *Statistics & Probability Letters* **8** 97–107.

[4] 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.

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

[6] Brown, B. M. (1983). Statistical use of spatial median. *J. R. Statist. Soc.* B **45**, 25–30.

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

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

[9] Chakraborty, B. (1996). On multivariate median regression. Technical Report No. 9/96, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, Calcutta.

[10] Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. In *Proc. of the American Mathematical Society*, Vol. 124, pp. 2539–2547.

[11] Chakraborty, B. and Chaudhuri, P. (1997). On an adaptive transformation and retransformation estimate of multivariate location. *J. R. Statist. Soc.* B. To appear.

[12] Chakraborty, B., Chaudhuri, P. and Oja, H. (1997). Operating transformation retransformation on spatial median and angle test. *Statistica Sinica.* To appear.

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

[14] 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, pp. 169-185. Amsterdam: North Holland.

[15] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Am. Statist. Assoc.* **91**, 862–872.

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

[17] Davis, J. B. and McKean, J. W. (1993). Rank based methods for multivariate linear models. *J. Am. Statist. Assoc.* **88**, 245–251.

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

[19] 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.

[20] Hettmansperger, T. P. and McKean, J. W. (1978). Statistical inference based on ranks. *Psychometrika* **43**, 69–79.

[21] Hettmansperger, T. P. and McKean, J. W. (1983). A geometric interpretation of inferences based on ranks in the linear model. *J. Am. Statist. Assoc.* **78**, 885–893.

[22] Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Statist.* **43**, 1449–1458.

[23] Jureckova, J. (1971). Nonparametric estimation of regression coefficients. *Ann. Math. Statist.* **42**, 1328–1338.

[24] Jureckova, J. (1973). Central limit theorem for Wilcoxon rank statistics process. *Ann. Statist.* **1**, 1046–1060.

[25] Koenker, R. and Portnoy, S. (1990). *M*-estimation of multivariate regression. *J. Am. Statist. Assoc.* **85**, 1060–1068.

[26] Koul, H. L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40**, 1950–1979.

[27] Koul, H. L. (1970). A class of ADF tests for subhypotheses in multiple linear regression. *Ann. Math. Statist.* **41**, 1273–1281.

[28] Lehmann, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34**, 957–966.

[29] Lehmann, E. L. (1963b). Asymptotically nonparametric inference : an alternative approach to linear models. *Ann. Math. Statist.* **34**, 1494–1506.

[30] Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Statist.* **35**, 726–734.

[31] Mottonen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparam. Statist.* **5**, 201–213.

[32] Puri, M. L. and Sen, P. K. (1969). A class of rank order tests for a general linear hypothesis. *Ann. Math. Statist.* **40**, 1325–1343.

[33] Puri, M. L. and Sen, P. K. (1973). A note on asymptotically distribution free tests for subhypotheses in multiple linear regression. *Ann. Statist.* **1**, 553–556.

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

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