JOURNAL OF MULTIVARIATE ANALYSIS **63**, 119–135 (1997)
ARTICLE NO. MV971694

General *M*-Estimation

Z. D. Bai

Department of Applied Mathematics, National Sun Yat-sen University,
Kaohsiung, Taiwan

and

Y. Wu

Department of Mathematics and Statistics, York University,
4700 Keele Street, North York, Ontario, Canada M3J 1P3

In this paper, a general form of *M*-estimation is proposed and some asymptotics are investigated. The model covers all linear and nonlinear regression models, AR time series, EIVR models, etc. as its special cases. The dispersion functions may be convex or differences of convex functions, the later covers almost all useful choices of the dispersion functions. © 1997 Academic Press

1. INTRODUCTION

In the past three decades, there are considerable works in the literature devoted to developing statistical procedures that are resistant to outliers and stable (or robust) with respect to deviations from a given distributional model. In particular, methods for robust regression, estimation, and testing on regression models have received much attention. Among these, procedures based on *M*-estimators play an important and complementary role. Reference may be made to papers by Huber (1964, 1967, 1973, 1981), Bickel (1975), Yohai and Maronna (1979), Heiler and Willers (1988), Basawa and Koul (1988), Bai, Rao and Wu (1992) and Bai, Rao and Wu (1997) among others.

In linear models, the regressors are assumed to be linear functions of the regression coefficients. This assumption may be due to mathematical convenience in obtaining the estimates by the traditional Least Squares (LS) method. For relaxing this probably restrictive assumption and for seeking

Received January 3, 1996; revised May 13, 1997.

AMS 1980 subject classification: 62J05; 62H12.

Key words and phrases: General *M*-estimation, asymptotics, multivariate linear model.

0047-259X/97 $25.00

robustness, Huber proposed the well known M-estimation. In M-estimation, for a prechosen dispersion function ρ, we are considering loss functions $\rho(\mathbf{y}_i - X_i\boldsymbol{\beta})$. In this case, the mathematical advantage of linear regressors disappears, but the difficulty is overcome by the modern computing techniques. Therefore, comparing efforts in computing the M-estimators, the same amount of effort will be needed for a more general model of nonlinear regression, i.e., to consider the target functions $\rho(\mathbf{y}_i - g(X_i, \boldsymbol{\beta}))$, as that for linear models. In usual regression models (linear or nonlinear), the distributions of the errors are assumed to be independent of the regression coefficients $\boldsymbol{\beta}$. However, in many practical situations, the distributions of the errors may depend on the regressors. For example, the distribution of the observation y_i is log-normally distributed with a mean value $\mathbf{x}_i'\boldsymbol{\beta}$. Then, the distribution of the error $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ will depend upon $\mathbf{x}_i'\boldsymbol{\beta}$. Because we only need to consider the function $\rho(\mathbf{y}_i - g(X_i, \boldsymbol{\beta}))$ where the form of $g(X_i, \boldsymbol{\beta})$ is not essential, we may simply consider a more general form of $\rho_i(\mathbf{y}_i, \boldsymbol{\beta})$. Therefore, in this paper we shall introduce a more general set-up for M-estimation, which will cover all the above mentioned models as its special cases.

Let $\{\mathbf{y}_1, ..., \mathbf{y}_n, ...\}$ be a sequence of random vectors and for each $\boldsymbol{\beta} \in \Omega$, $\{\rho_1(\mathbf{y}_1, \boldsymbol{\beta}), ..., \rho_n(\mathbf{y}_n, \boldsymbol{\beta}), ...\}$ be a sequence of functions which are differentiable about $\boldsymbol{\beta}$ for almost all $\mathbf{y}$'s, where Ω is an open convex subset of R^p known as the parameter space. Then the general M-estimate $\hat{\boldsymbol{\beta}}$ is defined as any value of $\boldsymbol{\beta}$ minimizing

$$\sum_{i=1}^{n} \rho_i(\mathbf{y}_i, \boldsymbol{\beta}). \tag{1.1}$$

Let $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta})$ denote the derivative of $\rho_i(\mathbf{y}_i, \boldsymbol{\beta})$ about $\boldsymbol{\beta}$ if the derivative exists and 0 otherwise. Then, $\hat{\boldsymbol{\beta}}$ satisfies

$$\sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \hat{\boldsymbol{\beta}}) = 0, \tag{1.2}$$

if the left side of (1.2) is continuous at $\hat{\boldsymbol{\beta}}$, or

$$\left| \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \hat{\boldsymbol{\beta}}) \right| = \min_{\boldsymbol{\beta}} \left| \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) \right|, \tag{1.3}$$

otherwise.

To investigate the asymptotics of the estimator, we need to define the "true parameter." Assume that there exists a vector $\boldsymbol{\beta}_0 \in \Omega$ and for each i there are an nonnegative definite matrix G_i and a function $\boldsymbol{\eta}_i(\boldsymbol{\beta})$ such that

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = G_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{\eta}_i(\boldsymbol{\beta}) \tag{1.4}$$

with $\boldsymbol{\eta}_i(\boldsymbol{\beta}) = o(Q_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0))$ as $\boldsymbol{\beta} \to \boldsymbol{\beta}_0$, where Q_i is some nonnegative definite matrix.

Although $\boldsymbol{\beta}_0$ may not be uniquely determined by a single Eq. (1.4) for a fixed i, in practice, one may show that $\boldsymbol{\beta}_0$ can be uniquely determined by all equations of type (1.4) for $i = 1, 2, ..., n$ $(n \geqslant n_0)$, under certain reasonable conditions.

Comparing the setup described above with that in a fundamental work of Huber (1967), one finds that our model generalizes Huber's, except that Huber defines his estimator $T(x)$ as an approximate minimizer of his target function instead of as the exact one, (see his formula (1)). This assumption is somewhat more realistic, since the recursive computation can only get an approximation of the exact M-estimator, however, there are no any theoretical difference in their asymptotic theory.

In Section 2, we shall give some further assumptions and state and establish the asymptotic results of the general M-estimation under convex discrepancy functions. In Section 3, the main results of Section 2 will be generalized to the general M-estimation under general discrepancy functions. Several examples and some discussions will be given in Section 4.

2. ASYMPTOTICS FOR CONVEX REGRESSION

In this section, we assume that $\rho_i(\mathbf{y}_i, \boldsymbol{\beta})$ is a convex function when $\mathbf{y}_i$ is fixed, $i = 1, 2, ...$ and denote by $\boldsymbol{\psi}_i$ a measurable selection of subgradients of ρ_i. We also assume that only on a set of probability zero $\boldsymbol{\psi}_i$ may have discontinuities.

Let $D_i(\mathbf{y}_i, \boldsymbol{\beta}) = \rho_i(\mathbf{y}_i, \boldsymbol{\beta} + \boldsymbol{\beta}_0) - \rho_i(\mathbf{y}_i, \boldsymbol{\beta}_0)$, $\boldsymbol{\Delta}_i(\mathbf{y}_i, \boldsymbol{\beta}) = \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta} + \boldsymbol{\beta}_0) - \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)$ and $\boldsymbol{\Delta} = \sum_{i=1}^n \boldsymbol{\Delta}_i$, $\bar{Q}(n) = \sum_{i=1}^n Q_i$, $G(n) = \sum_{i=1}^n G_i$, and $S(\boldsymbol{\beta}, A) = \boldsymbol{\beta}' A \boldsymbol{\beta}$. For convenience of notation, we shall write $\boldsymbol{\Delta}_i$, $\bar{Q}$ and G for $\boldsymbol{\Delta}_i(\mathbf{y}_i, \boldsymbol{\beta})$, $\bar{Q}(n)$ and $G(n)$, respectively, when there is no confusion.

We make the following assumptions.

(A1) Let $Q = Q(n)$ be a positive definite matrix and suppose that $0 \leqslant a_1 < \inf \lambda_{\min}(G^Q) \leqslant \sup \lambda_{\max}(G^Q) < a_2 < \infty$, and that $0 \leqslant a_1 < \inf \lambda_{\min}(\bar{Q} Q^{-1}) \leqslant \sup \lambda_{\max}(\bar{Q} Q^{-1}) < a_2 < \infty$, where, and in what follows, $G^Q = Q^{-1/2} G Q^{-1/2}$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A, respectively. $Q^{-1} \to 0$.

Remark. In general, Q can be chosen as $Q = \mathrm{E}(\sum_{i=1}^n \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0))(\sum_{i=1}^n \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0))'$ and $Q_i = \mathrm{E} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0) \boldsymbol{\psi}_i'(\mathbf{y}_i, \boldsymbol{\beta}_0)$. Then $\bar{Q} = Q$ when $\mathbf{y}_1, \mathbf{y}_2, ...$ are independent.

(A2) $\operatorname{Cov}(\bar{D}_i(\mathbf{y}_i, \boldsymbol{\beta}), \bar{D}_j(\mathbf{y}_j, \boldsymbol{\beta})) = v_{i,j}(\boldsymbol{\beta})\, \boldsymbol{\beta}'(Q_i + Q_j)\, \boldsymbol{\beta}$ and $\max_i \sum_{j=1}^n |v_{i,j}(\boldsymbol{\beta})| \to 0$, as $|\boldsymbol{\beta}| \to 0$, where

$$\bar{D}_i(\mathbf{y}_i, \boldsymbol{\beta}) = D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta}'\psi_i(\mathbf{y}_i, \boldsymbol{\beta}_0) - \mathrm{E}(D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta}'\psi_i(\mathbf{y}_i, \boldsymbol{\beta}_0)).$$

Remark. We point out that the second part of Condition (A2) is a consequence if $\mathrm{E}\, \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i' \leqslant v_i(\boldsymbol{\beta})\, Q_i$ with $\max_{i \leqslant n} v_i(\boldsymbol{\beta}) \to 0$ (as $|\boldsymbol{\beta}| \to 0$) and the sequence $\{\mathbf{y}_i\}$ is ϕ-mixing with $\sum \phi_n^{1/2} < \infty$. In fact, by Lemma 1 of Section 20 in Billingsley (1968), we have

$$|\operatorname{Cov}(\bar{D}_i(\mathbf{y}_i, \boldsymbol{\beta}), \bar{D}_j(\mathbf{y}_i, \boldsymbol{\beta}))| \leqslant 2\phi_{|i-j|}^{1/2}[\operatorname{Var}(\bar{D}_i(\mathbf{y}_i, \boldsymbol{\beta}))\operatorname{Var}(\bar{D}_j(\mathbf{y}_j, \boldsymbol{\beta}))]^{1/2}.$$

Then by the convexity of ρ_i, we have

$$|D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta}'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)| \leqslant |\boldsymbol{\beta}'\boldsymbol{\Delta}_i|.$$

Hence,

$$\operatorname{Var}(\bar{D}_i(\mathbf{y}_i, \boldsymbol{\beta})) \leqslant \boldsymbol{\beta}'\mathrm{E}\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i'\boldsymbol{\beta}.$$

Then, Condition (A2) is true by choosing $v_{ij} \leqslant 2\sqrt{|\phi_{|i-j|} v_i v_j|}$.

(A3) $Q(n)^{-1/2} \sum_{i=1}^n \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N(0, I_p)$.

Remark. If $\mathbf{y}_i$'s are independent, then $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)$'s are independent. Hence Assumption (A3) is true under Lindeberg's condition. In many applications of the main theorems, the independence condition holds.

We have the following theorems.

THEOREM 2.1. *Under the assumptions* (A1) *and* (A2), *for any fixed* $\mu > 0$,

$$\sup_{|Q^{1/2}\boldsymbol{\beta}| < \mu} \left| \sum_{i=1}^n [D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta}'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] - \tfrac{1}{2}S(\boldsymbol{\beta}, G) \right| \to 0 \qquad \textit{in probability}. \tag{2.1}$$

The following lemma is needed in the proof of Theorem 2.1.

LEMMA 2.1. *We have*

$$\mathrm{E}D_i(\mathbf{y}_i, \boldsymbol{\beta}) = \tfrac{1}{2}S(\boldsymbol{\beta}, G_i) + o(S(\boldsymbol{\beta}, Q_i)). \tag{2.2}$$

Proof. The lemma can be proved following the same procedure as the proof of Lemma 1 in Bai, Rao and Wu (1992).

Proof of Theorem 2.1. In order to prove (2.1), we only need to show that for any subsequence $\{n'\}$ of positive integers, there exists a subsequence $\{n''\}$ of the subsequence $\{n'\}$ such that

$$\sup_{|Q^{1/2}\boldsymbol{\beta}|<\mu}\left|\sum_{i=1}^{n''}[D_i(\mathbf{y}_i,\boldsymbol{\beta})-\boldsymbol{\beta}'\boldsymbol{\psi}_i(\mathbf{y}_i,\boldsymbol{\beta}_0)] - \tfrac{1}{2}S(\boldsymbol{\beta},G)\right|\to 0 \qquad \text{a.s. as} \quad n''\to\infty. \tag{2.3}$$

Let us make the transformation $\boldsymbol{\gamma}=Q^{1/2}\boldsymbol{\beta}$. Then, the assertion (2.1) becomes

$$\sup_{|\boldsymbol{\gamma}|<\mu}\left|\sum_{i=1}^{n''}[D_i(\mathbf{y}_i,Q^{-1/2}\boldsymbol{\gamma})-\boldsymbol{\gamma}'Q^{-1/2}\boldsymbol{\psi}_i(\mathbf{y}_i,\boldsymbol{\beta}_0)] - \tfrac{1}{2}S(Q^{-1/2}\boldsymbol{\gamma},G)\right|\to 0 \qquad \text{a.s. as} \quad n''\to\infty. \tag{2.4}$$

By (A2) and (A3), we have

$$\begin{aligned}
\operatorname{Var}&\left(\sum_{i=1}^{n}(D_i(\mathbf{y}_i,Q^{-1/2}\boldsymbol{\gamma})-\boldsymbol{\gamma}'Q^{-1/2}\boldsymbol{\psi}_i(\mathbf{y}_i,\boldsymbol{\beta}_0))\right)\\
&=\sum_{i,j}\operatorname{Cov}(\bar{D}_i(\mathbf{y}_i,\boldsymbol{\beta}),\bar{D}_j(\mathbf{y}_j,\boldsymbol{\beta}))\\
&\leqslant 2\sum_{i=1}^{n}\boldsymbol{\beta}'Q_i\boldsymbol{\beta}\sum_{j=1}^{n}|v_{ij}(\boldsymbol{\beta})|\\
&\leqslant 2\boldsymbol{\gamma}'Q^{-1/2}\bar{Q}Q^{-1/2}\boldsymbol{\gamma}\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|v_{ij}(Q^{-1/2}\boldsymbol{\gamma})|\\
&\leqslant 2a_1^{-1}|\boldsymbol{\gamma}|^2\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|v_{ij}(Q^{-1/2}\boldsymbol{\gamma})|\to 0, \qquad \text{since} \quad Q^{-1}\to 0.
\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}[D_i(\mathbf{y}_i,Q^{-1/2}\boldsymbol{\gamma})-\boldsymbol{\gamma}'Q^{-1/2}\boldsymbol{\psi}_i(\mathbf{y}_i,\boldsymbol{\beta}_0) - \mathrm{E}D_i(\mathbf{y}_i,Q^{-1/2}\boldsymbol{\gamma})]\to 0 \qquad \text{in probability.}$$

By Lemma 2.1,

$$\begin{aligned}
\sum_{i=1}^{n}\mathrm{E}D_i(\mathbf{y}_i,Q^{-1/2}\boldsymbol{\gamma})&=\tfrac{1}{2}S(Q^{-1/2}\boldsymbol{\gamma},G)+o(S(Q^{-1/2}\boldsymbol{\gamma},G))\\
&=\tfrac{1}{2}S(\boldsymbol{\gamma},G^{Q})+o(1),
\end{aligned}$$

where G^Q is defined in Assumption (A1). By (A1), there is a subsequence $\{n'''\}$ of $\{n'\}$ such that

$$G^Q(n''') \to G_0^Q > 0 \qquad \text{as} \quad n_{ij} \to \infty,$$

where G_0^Q is some positive definite matrix of constants. Then, for all fixed $\boldsymbol{\gamma}$, we have

$$\sum_{i=1}^{n'''} [D_i(\mathbf{y}_i, Q^{-1/2}\boldsymbol{\gamma}) - \boldsymbol{\gamma}' Q^{-1/2} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)]$$
$$- \tfrac{1}{2} S(\boldsymbol{\gamma}, G_0^Q) \to 0 \qquad \text{in probability as} \quad n''' \to \infty.$$

Using diagonalization technique, we can choose a subsequence $\{n''\}$ of $\{n'''\}$ such that

$$\sum_{i=1}^{n''} [D_i(\mathbf{y}_i, Q^{-1/2}\boldsymbol{\gamma}) - \boldsymbol{\gamma}' Q^{-1/2} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)]$$
$$\to \tfrac{1}{2} S(\boldsymbol{\gamma}, G_0^Q) \to 0, \qquad \text{a.s. as} \quad n'' \to \infty,$$

for each $\boldsymbol{\gamma}$ in any given countable dense set of R^p. Since $\sum_{i=1}^{n''} [D_i(\mathbf{y}_i, Q^{-1/2}\boldsymbol{\gamma}) - \boldsymbol{\gamma}' Q^{-1/2} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)]$ is convex in $\boldsymbol{\gamma}$ and $\frac{1}{2} S(\boldsymbol{\gamma}, G_0^Q)$ is a continuously differentiable convex function in $\boldsymbol{\gamma}$, we obtain, by the generalized Theorem 10.8 of Rockafellar (1970),

$$\sup_{|Q^{1/2}\boldsymbol{\beta}| < \mu} \left| \sum_{i=1}^{n''} [D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] - \tfrac{1}{2} S(\boldsymbol{\gamma}, G_0^Q) \right| \to 0 \qquad \text{a.s.}$$

Therefore, (2.4) follows and the proof of Theorem 2.1 is complete.

THEOREM 2.2. *In addition to the assumptions of Theorem* 2.1, *we assume the constant* $a_1 > 0$ *in the condition* (A1), *then*

$$\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0 \qquad \textit{in probability.} \tag{2.5}$$

THEOREM 2.3. *Under the assumptions of Theorem* 2.2, *we have*, *for any* $\mu > 0$,

$$\sup_{|Q^{1/2}\boldsymbol{\beta}| < \mu} |Q^{-1/2}(\boldsymbol{\Delta} - G\boldsymbol{\beta})| \to 0 \qquad \textit{in probability.} \tag{2.6}$$

Based on Theorem 2.1, Theorems 2.2 and 2.3 can be proved by the same approach of Bai, Rao and Wu (1992).

THEOREM 2.4. *Under the assumptions of Theorem 2.2, we have*

$$\hat{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}} + o_p(\|Q^{-1/2}\|), \tag{2.7}$$

where

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + G^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0). \tag{2.8}$$

Consequently, if we further assume (A3) is true, then

$$Q^{-1/2}G(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, I). \tag{2.9}$$

Proof. In order to prove (2.7), we need only to show that

$$Q^{1/2}(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \to 0 \qquad \text{in probability.} \tag{2.10}$$

Then (2.9) follows when (A3) is true.

By (A1), we have

$$Q^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(1).$$

Hence, by (2.1) and the definition of $\bar{\boldsymbol{\beta}}$, it follows that

$$\sum_{i=1}^{n} D_i(\mathbf{y}_i, \bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \tfrac{1}{2} S(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, G) \to 0 \qquad \text{in probability.} \tag{2.11}$$

Take $\tau > 0$. For a sequence $\{\mu_n\}$ with $\mu_n \to \infty$, by (2.1), we have

$$\sup_{|Q^{1/2}\boldsymbol{\beta}| \leqslant \mu_n + \tau} \left| \sum_{i=1}^{n} [D_i(\mathbf{y}_i, \boldsymbol{\beta}) - \boldsymbol{\beta}'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] - \tfrac{1}{2} S(\boldsymbol{\beta}, G) \right| \to 0 \quad \text{in probability,}$$

which, together with (2.11), implies that

$$\sup_{|Q^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})| = \tau} \left| \sum_{i=1}^{n} [\rho_i(\mathbf{y}_i, \boldsymbol{\beta}) - \rho_i(\mathbf{y}_i, \bar{\boldsymbol{\beta}}) - \tfrac{1}{2} S(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}, G)) \right| \to 0 \qquad \text{in probability.} \tag{2.12}$$

Since for $|Q^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})| = \tau$,

$$S(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}, G) \geqslant \lambda_{\min}(G^Q)\, \tau^2,$$

by (2.12), we get

$$P(|Q^{1/2}(\hat{\boldsymbol\beta}-\bar{\boldsymbol\beta})|\geqslant\tau)\to 0.$$

Then, (2.10) and hence the theorem is proved.

Now, consider a test of the hypothesis H_0: $H'\boldsymbol\beta=\xi_0$, where H is a $p\times q$ matrix of rank q. Let $\tilde{\boldsymbol\beta}$ denote the solution of

$$\min_{H'\boldsymbol\beta=\xi_0}\sum_{i=1}^{n}\rho_i(\mathbf{y}_i,\boldsymbol\beta)$$

and $\hat{\boldsymbol\beta}$ be defined as before. Then we have the following theorem.

THEOREM 2.5. *Suppose the assumption* (A3) *holds in addition to the assumptions of Theorem 2.2. Then, under the null hypothesis,*

$$\left|\sum_{i=1}^{n}[\rho_i(\mathbf{y}_i,\tilde{\boldsymbol\beta})-\rho_i(\mathbf{y}_i,\hat{\boldsymbol\beta})]-\tfrac{1}{2}\left|K'\sum_{i=1}^{n}\boldsymbol\psi_i(\mathbf{y}_i,\boldsymbol\beta_0)\right|^2\right|\to 0\qquad\textit{in probability,}\tag{2.13}$$

and

$$(H'\hat{\boldsymbol\beta}-\xi_0)'\,(H'G^{-1}QG^{-1}H)^{-1}\,(H'\hat{\boldsymbol\beta}-\xi_0)\to\chi_q^2,\tag{2.14}$$

where

$$K=G^{-1}H(H'G^{-1}H)^{-1/2},$$

is a $p\times q$ *matrix and* χ_q^2 *denotes a chi-square random variable with* q *degrees of freedom.*

Proof. By (2.8) and (2.10), it follows that

$$Q^{1/2}(\hat{\boldsymbol\beta}-\boldsymbol\beta_0)=Q^{1/2}G^{-1}\sum_{i=1}^{n}\boldsymbol\psi_i(\mathbf{y}_i,\boldsymbol\beta_0)+o_p(1).\tag{2.15}$$

By (2.1), we get

$$\sum_{i=1}^{n}[\rho_i(\mathbf{y}_i,\hat{\boldsymbol\beta})-\rho_i(\mathbf{y}_i,\boldsymbol\beta_0)]=-\tfrac{1}{2}\left|G^{-1/2}\sum_{i=1}^{n}\boldsymbol\psi_i(\mathbf{y}_i,\boldsymbol\beta_0)\right|^2+o_p(1).\tag{2.16}$$

Let W be a $p\times(p-q)$ matrix such that $K'W=0$ and $W'W=I_{p-q}$ (with two extreme cases where we define $K=0$, $W=I_p$ if $q=0$ and $K=I_p$, $W=0$ if $p=q$). Then

$$H'(\boldsymbol\beta-\boldsymbol\beta_0)=0\Leftrightarrow\boldsymbol\beta=\boldsymbol\beta_0+W\boldsymbol\alpha.$$

That means, under null hypothesis, we may assume

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + W\boldsymbol{\alpha}.$$

Let $\hat{\boldsymbol{\alpha}}$ be the *M*-estimator of $\boldsymbol{\alpha}$ with respect to the dispersion functions $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}_0 + W\boldsymbol{\alpha})$. Note that the gradient of $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}_0 + W\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ is $W'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0 + W\boldsymbol{\alpha})$ with

$$\mathrm{E}\, W'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0 + W\boldsymbol{\alpha}) = W'G_i W(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + o(W'Q_i W(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)).$$

Let $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + W\hat{\boldsymbol{\alpha}}$. Similar to (2.16), we obtain

$$\sum_{i=1}^{n} [\rho_i(\mathbf{y}_i, \tilde{\boldsymbol{\beta}}) - \rho_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] = -\tfrac{1}{2} \left| (W'GW)^{-1/2}\, W' \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0) \right|^2 + o_p(1),$$

under the null hypothesis. Hence,

$$\sum_{i=1}^{n} [\rho_i(\mathbf{y}_i, \tilde{\boldsymbol{\beta}}) - \rho_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] = \tfrac{1}{2} \left| K' \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0) \right|^2 + o_p(1), \qquad (2.17)$$

by noticing that

$$\begin{aligned} KK' &= G^{-1} - W(W'GW)^{-1}\, W' \\ &= G^{-1}H(H'G^{-1}H)^{-1}\, H'G^{-1}. \end{aligned}$$

By Theorem 2.4 and (2.15), it follows that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\, H(H'G^{-1}QG^{-1}H)^{-1}\, H'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \chi_q^2.$$

The proof of the theorem is finished.

3. COMMENTS ON NON-CONVEX DISPERSIONS

In many situations of robust estimation, the dispersion functions may not be convex. However, the results in Section 2 can be easily extended to the case that each dispersion function is a difference of two convex functions, which covers almost all useful cases of *M*-estimation by the following fact.

THEOREM 3.1 (See Theorem 4.2 of Bai, Rao and Wu (1997)). *Every function which is continuously twice differentiable can be written as a difference of two convex functions which are strictly convex.*

Therefore, we shall devote this section to some general comments on *M*-estimation defined by dispersion functions each of which is a difference of two convex functions. We shall only state the main results without detailed mathematics.

To begin with, we need to clarify the following concept of the M-estimators. When ρ_i is not convex, the global minimizer of (1.1) may not exist or may not be consistent even when it exists. See Bai, Rao and Wu (1997) for examples. However, it can be shown that the minimization problem (1.1) must have at least one local minimizer around the true value of the regression parameter even when the global minimizer does not exist. At the same time, there arises another problem that (1.1) may have many local minimizers in such case. Therefore, we have to clarify which minimizer satisfies the asymptotic properties discussed in this section.

For this end, let δ be a positive constant and denote by $\hat{\boldsymbol{\beta}}(\delta)$ the absolute minimizer of (1.1) in the neighborhood $\{\boldsymbol{\beta}: |Q^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)| \leqslant \delta\}$, where Q is defined in (B1). In case that the solution is not unique, $\hat{\boldsymbol{\beta}}(\delta)$ denotes either one of the solutions. If there is a sequence $\{\delta_n: \delta_n \to \infty\}$ such that $Q^{1/2}(\hat{\boldsymbol{\beta}}(\delta_n)-\boldsymbol{\beta}_0)=O(1)$, then we denote $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(\delta_n)$. It should be noted that the definition of $\hat{\boldsymbol{\beta}}$ is actually independent of the choice of the sequence $\{\delta_n\}$ because of the fact that $|Q^{1/2}(\hat{\boldsymbol{\beta}}(\delta_2)-\boldsymbol{\beta}_0)| \leqslant \delta_1 \leqslant \delta_2$ implies that $\hat{\boldsymbol{\beta}}(\delta_1)=\hat{\boldsymbol{\beta}}(\delta_2)$. In the present section, we shall show that such a sequence always exists.

Let $\rho_i(\mathbf{y}_i, \boldsymbol{\beta})=\rho_{i1}(\mathbf{y}_i, \boldsymbol{\beta})-\rho_{i2}(\mathbf{y}_i, \boldsymbol{\beta})$, where ρ_{i1}, ρ_{i2} are strictly convex as functions of $\boldsymbol{\beta}$ with fixed $\mathbf{y}_i$, and have derivatives $\boldsymbol{\psi}_{i1}(\mathbf{y}_i, \boldsymbol{\beta})$ and $\boldsymbol{\psi}_{i2}(\mathbf{y}_i, \boldsymbol{\beta})$ about $\boldsymbol{\beta}$, respectively, for $i=1, 2, ...$ It is assumed that the union of discontinuity set of $\boldsymbol{\psi}_{i1}, \boldsymbol{\psi}_{i2}$, $i=1, 2, ...,$ has zero probability and $\mathrm{E}\boldsymbol{\psi}_{ij}(\mathbf{y}_i, \boldsymbol{\beta})=G_{ij}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\boldsymbol{\eta}_{ij}(\boldsymbol{\beta})$, with $G_{ij} \geqslant 0$ and $\boldsymbol{\eta}_{ij}(\boldsymbol{\beta})=o(Q_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0))$ as $\boldsymbol{\beta} \to \boldsymbol{\beta}_0$, $j=1, 2$.

As in Section 2, for each $j=1$ and 2, define $D_{ij}=D_{ij}(\mathbf{y}_i, \boldsymbol{\beta})=\rho_{ij}(\mathbf{y}_i, \boldsymbol{\beta}+\boldsymbol{\beta}_0)-\rho_{ij}(\mathbf{y}_i, \boldsymbol{\beta}_0)$, $\boldsymbol{\Delta}_{ij}=\boldsymbol{\Delta}_{ij}(\mathbf{y}_i, \boldsymbol{\beta})=\boldsymbol{\psi}_{ij}(\mathbf{y}_i, \boldsymbol{\beta}+\boldsymbol{\beta}_0)-\boldsymbol{\psi}_{ij}(\mathbf{y}_i, \boldsymbol{\beta}_0)$, $\boldsymbol{\Delta}(j)=\sum_{i=1}^n \boldsymbol{\Delta}_{ij}$, and $G^j=G^j(n)=\sum_{i=1}^n G_{ij}$. And denote $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta})=\boldsymbol{\psi}_{i1}(\mathbf{y}_i, \boldsymbol{\beta})-\boldsymbol{\psi}_i(\mathbf{y}_{i2}, \boldsymbol{\beta})$, $\boldsymbol{\Delta}=\boldsymbol{\Delta}(1)-\boldsymbol{\Delta}(2)$ and $G=G(n)=G^1-G^2$.

We make the following assumptions.

(B1) Suppose that G satisfies the Assumption (A1) with $a_1>0$ and G^2 satisfies (A1) with $a_1 \geqslant 0$, where $G=G^1-G^2$.

(B2) The Assumption (A2) is true for both ρ_{i1} and ρ_{i2}.

(B3) Same as (A3).

THEOREM 3.2. *Under the assumptions* (B1) *and* (B2), *for any fixed* $\mu>0$,

$$\sup_{|Q^{1/2}\boldsymbol{\beta}|<\mu} \left| \sum_{i=1}^{n} [D_i(\mathbf{y}_i, \boldsymbol{\beta})-\boldsymbol{\beta}'\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)] - \tfrac{1}{2} S(\boldsymbol{\beta}, G) \right| \to 0 \qquad \textit{in probability.} \tag{3.1}$$

Proof. By Theorem 2.1, both D_{i1} associated with $\boldsymbol{\psi}_{i1}(\mathbf{y}_i, \boldsymbol{\beta}_0)$ and D_{i2} associated with $\boldsymbol{\psi}_{i2}(\mathbf{y}_i, \boldsymbol{\beta}_0)$ satisfy (2.1). Thus, (3.1) follows.

THEOREM 3.3. *Under the assumptions of Theorem* 3.2, *we have a local minimizer* $\hat{\boldsymbol{\beta}}$ *such that*

$$\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0 \qquad \textit{in probability.}$$

The proof of this theorem is completely the same as that of Theorem 2.2. Here, we would like to remind the reader that the local absolute minimizer is isolated in a small ball with center $\bar{\boldsymbol{\beta}}$, where $\bar{\boldsymbol{\beta}}$ is similarly defined as in (2.8). Concretely speaking, we have the following:

Remark. Since (3.1) is true for all $\mu > 0$, there exists a sequence of $\mu_n \to \infty$ such that (3.1) is still true when μ is replaced by μ_n. Using similar arguments as in the proof of Theorem 2.2, one can prove that with a probability tending to one, there is no local minimizer in the hyper ring $0 < \delta < |\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}| < \mu_n$.

THEOREM 3.4. *Under the assumptions of Theorem* 3.2, *we have*, *for any* $\mu > 0$,

$$\sup_{|Q^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)| < \mu} |Q^{-1/2}(\boldsymbol{\Delta} - G(\boldsymbol{\beta} - \boldsymbol{\beta}_0))| \to 0 \qquad \textit{in probability.}$$

This theorem follows from the fact that the above convergence is true for both $\boldsymbol{\Delta}(1)$ and $\boldsymbol{\Delta}(2)$.

THEOREM 3.5. *Under the assumptions* (B1), (B2) *and* (B3), *we have*

$$Q^{-1/2}G(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, I).$$

For testing the hypothesis H_0: $H'\boldsymbol{\beta} = \xi_0$, where H is a $p \times q$ matrix of rank q. Let $\tilde{\boldsymbol{\beta}}$ denote the solution of

$$\min_{H'\boldsymbol{\beta} = \xi_0} \sum_{i=1}^{n} \rho_i(\mathbf{y}_i, \boldsymbol{\beta})$$

and $\hat{\boldsymbol{\beta}}$ be defined as before. Then, under the assumptions (B1), (B2) and (B3), Theorem 2.5 is still true when the dispersion functions ρ_i's are replaced by differences of convex functions. Since its mathematical expressions are exactly the same as those in Theorem 2.5, we do not write it out as a theorem.

4. SOME EXAMPLES

In this section, we implicitly assume that ρ or ρ_i are convex functions or differences of convex functions without stating. For most well known results in the literature, Assumptions (A1–3) or (B1–3) are trivially satisfied without verification, which is even not clearly stated here. We only give a bit more details to cases which to not follow from the usual M-estimation.

4.1. *The usualM-estimation in linear regression models.*

Consider a general multivariate regression model

$$\mathbf{y}_i = X_i'\boldsymbol{\beta} + \mathbf{e}_i,\ i=1, ..., n \tag{4.1}$$

where $\mathbf{e}_i$, $i=1, ..., n$, are vectors of random errors, X_i, $i=1, ..., n$, are design matrices and $\boldsymbol{\beta}$ is a p-vector of unknown parameters.

In Bai, Rao and Wu (1992) and Bai, Rao and Wu (1997), a general asymptotic theory of M-estimation is developed under dispersions formed by a convex function and a difference of convex functions, respectively. That means, the M-estimator $\hat{\boldsymbol{\beta}}$ is defined by minimizing

$$\sum_{i=1}^{n} \rho(\mathbf{y}_i - X_i'\boldsymbol{\beta}) \tag{4.2}$$

for a suitable choice of the function ρ, or solving the equations

$$\sum_{i=1}^{n} X_i \boldsymbol{\psi}(\mathbf{y}_i - X_i'\boldsymbol{\beta}) = 0 \tag{4.3}$$

for a suitable choice of the $\boldsymbol{\psi}$ function. A well known example for the convex and robust choice of ψ is Huber's example. The following are some examples of non-convex dispersions:

1. $\psi(x) = 2x/(1+x^2)$. Then, we can choose $\psi_1(x) = 2x/(1+x^2)$ for $|x| \leqslant 1$ and $=\text{sign}(x)$ for $|x| > 1$, whereas $\psi_2(x) = 0$ for $|x| \leqslant 1$ and $=\text{sign}(x) - 2x/(1+x^2)$ for $|x| > 1$.

2. Hampel's ψ, i.e., for constants, $0 < a < b < c$, $\psi(x) = x$ for $|x| \leqslant a$, $=a\,\text{sign}(x)$ for $a < |x| \leqslant b$, $=a\,\text{sign}(x)(c-|x|)/(c-b)$ for $b < |x| \leqslant c$ and $=0$ otherwise. In this case, we can choose $\psi_1(x) = x$ for $|x| \leqslant a$ and $=a\,\text{sign}(x)$ for $|x| > a$ whereas $\psi_2(x) = 0$ for $|x| \leqslant b$, $=a\,\text{sign}(x)(|x|-b)/(c-b)$ for $b < |x| \leqslant c$ and $=a\,\text{sign}(x)$, otherwise.

For these choices, the G_i and Q_i are defined by the expectation of $\psi(y_i - \mathbf{x}_i'\boldsymbol{\beta})$ and the covariance of $\psi(y_i - \mathbf{x}_i'\boldsymbol{\beta}_0)$, for $i=1, ..., n$. Under very general and mild regularity conditions, the assumptions (A1)–(A3) or

(B1)–(B3) are satisfied. For details, see Bai, Rao and Wu (1992) and Bai, Rao and Wu (1997).

4.2. *The usual M-estimation in nonlinear regression models.*

Consider a general multivariate nonlinear regression model

$$\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\beta}) + \mathbf{e}_i, \ i = 1, ..., n \tag{4.4}$$

where $\mathbf{e}_i$, $i = 1, ..., n$, are vectors of random errors, $\mathbf{f}_i$, $i = 1, ..., n$, are given functions and $\boldsymbol{\beta}$ is a p-vector of unknown parameters. As a special case, $\mathbf{f}_i(\boldsymbol{\beta}) = \mathbf{f}(X_i, \boldsymbol{\beta})$ with given $\mathbf{f}$ and design matrices X_i, $i = 1, ..., n$.

In M-estimation of $\boldsymbol{\beta}$, one takes $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}) = \rho(\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}))$ and $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = -(d\mathbf{f}_i^t/d\boldsymbol{\beta})\, \boldsymbol{\psi}(\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}))$.

Under the conditions that $\mathrm{E}(\boldsymbol{\psi}(\mathbf{e} + \mathbf{c})) = A\mathbf{c} + o(\mathbf{c})$ (or $= A\mathbf{c} + O(|\mathbf{c}|^2)$) with $A > 0$ and $\mathrm{E}(\boldsymbol{\psi}(\mathbf{e})\, \boldsymbol{\psi}(\mathbf{e})^t) = B > 0$ and some minor conditions, say, the derivatives $d\mathbf{f}_i/d\boldsymbol{\beta}$ are equi-continuous at $\boldsymbol{\beta}_0$, we have $G_i = (d\mathbf{f}_i^t/d\boldsymbol{\beta}) A (d\mathbf{f}_i/d\boldsymbol{\beta})|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0}$ and $Q_i = (d\mathbf{f}_i^t/d\boldsymbol{\beta}) B (d\mathbf{f}_i/d\boldsymbol{\beta})|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0}$. Furthermore, if some regularity conditions on the growth rate and nonsingularity conditions on $(d\mathbf{f}_i/d\boldsymbol{\beta})|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0}$ are met, say

$$\max_{1 \leqslant i \leqslant n} \frac{d\mathbf{f}_i^t}{d\boldsymbol{\beta}} \left(\sum_{j=1}^{n} \frac{d\mathbf{f}_j}{d\boldsymbol{\beta}} \frac{d\mathbf{f}_j^t}{d\boldsymbol{\beta}} \right)^{-1} \frac{d\mathbf{f}_i}{d\boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \to 0,$$

one can verify that the conditions of Sections 2 or 3 are satisfied.

4.3. *The general M-estimation in linear models.*

In linear models, the general M-estimation $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined by a value which minimizes

$$\sum_{i=1}^{n} w_i(X_i)\, \rho(\mathbf{y}_i - X_i \boldsymbol{\beta})$$

or a solution of the equation

$$\sum_{i=1}^{n} w_i(X_i)\, X_i \boldsymbol{\psi}(\mathbf{y}_i - X_i \boldsymbol{\beta}) = 0,$$

where $\{w_i(X_i)\}$ is a set of weights depending on the design. Then, this is equivalent to select $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}) = w_i(X_i)\, \rho(\mathbf{y}_i - X_i \boldsymbol{\beta})$ and $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = -w_i(X_i)\, X_i \boldsymbol{\psi}(\mathbf{y}_i - X_i \boldsymbol{\beta})$.

This is the well known Mallow estimation. For general results and the choices of $\{w_i(X_i)\}$, see Hampel *et al.* (1986).

4.4. *The M-estimation in AR model.*

Consider the AR time series

$$x_n = \beta_1 x_{n-1} + \cdots + \beta_p x_{n-p} + \varepsilon_n,$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a vector of unknown autoregressive coefficients and $\{\varepsilon_1, \varepsilon_2, ...\}$ is a sequence of iid. random errors. Then, M-estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined by choosing $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}) = \rho(x_i - \beta_1 x_{i-1} - \cdots - \beta_p x_{i-p})$, where ρ is a suitably chosen dispersion function and $\mathbf{y}_i = (x_i, ..., x_{i-p})'$ for $p < i \leqslant n$. If ψ is the derivative of ρ satisfying $\mathrm{E}\psi(\varepsilon + a) = \lambda a + o(a)$ and $\mathrm{E}\psi^2(\varepsilon) = \sigma^2$, then we have $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = -\mathbf{x}_i \psi(\varepsilon_i - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_0))$, where $\mathbf{x}_i = (x_{i-1}, ..., x_{i-p})'$.

Let $\Sigma = \mathrm{E}(\mathbf{x}_i \mathbf{x}_i')$. Then, one can verify that

$$\mathrm{E}(\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta})) = \lambda \Sigma(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(|\boldsymbol{\beta} - \boldsymbol{\beta}_0|) \tag{4.5}$$

and

$$\mathrm{E}(\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)\, \boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)') = \sigma^2 \Sigma. \tag{4.6}$$

Since an AR time series is ϕ-mixing whose ϕ function is exponentially decaying, by (4.5), (4.6) and the remark below Assumption (A2), Assumption (A2) is true. Hence, the conditions of Section 2 are true and consequently

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N(0, A),$$

where $A = \lambda^{-2} \sigma^2 \Sigma^{-1}$.

This has been studied by R. D. Martin (1980, 1981), to which the readers are referred for more details.

4.5. *The MLE of regression coefficients of means of exponential variables.*

Let y_i be exponentially distributed with a mean $\mathbf{x}_i' \boldsymbol{\beta}$, where $\mathbf{x}_i$ is known. We may write this model as

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where ε_i has a zero mean and variance $2(\mathbf{x}_i' \boldsymbol{\beta})^2$. That means, the error distributions depend on the unknown parameters.

Let $y_1, ..., y_n$ be independent. The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is to minimize

$$\sum_{i=1}^{n} \left(\frac{y_i}{\mathbf{x}_i' \boldsymbol{\beta}} + \log(\mathbf{x}_i' \boldsymbol{\beta}) \right),$$

under the restriction that $\min_i \mathbf{x}_i' \boldsymbol{\beta} > 0$.

Corresponding to our model, we have $\rho_i(\mathbf{y}_i, \boldsymbol{\beta}) = (y_i/\mathbf{x}_i'\boldsymbol{\beta}) + \log(\mathbf{x}_i'\boldsymbol{\beta})$, and then $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = -(\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})/(\mathbf{x}_i'\boldsymbol{\beta})^2)$. By elementary calculation, we obtain

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = \frac{\mathbf{x}_i\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{(\mathbf{x}_i'\boldsymbol{\beta})^2}$$

and

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)\, \boldsymbol{\psi}_i'(\mathbf{y}_i, \boldsymbol{\beta}_0) = 2\mathbf{x}_i\mathbf{x}_i'.$$

Then, it is easy to verify that the assumptions (A1–3) hold under the conditions

$$\inf_n \min_{i \leqslant n} \mathbf{x}_i'\boldsymbol{\beta}_0/|\mathbf{x}_i| \geqslant \delta > 0$$

and

$$\max_{i \leqslant n} \left[\mathbf{x}_i' \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' \right)^{-1} \mathbf{x}_i \left(\max\left(1, \frac{1}{\mathbf{x}_i'\boldsymbol{\beta}_0}\right)\right)\right] \to 0, \qquad \text{as} \quad n \to \infty.$$

Similar consideration may apply to other parametric models, such as inverse Gaussian regression (see Chhikara and Folks (1989)).

4.6. *The nonlinear regression in EIVR models.*

There is much work in Error In Variables Regression models. For an application to this model, we refer to Gleser (1990) and Stefanski (1985). The EIVR (structural) model is given by

$$\mathbf{y} = \mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon},\ \mathbf{x} = \boldsymbol{\xi} + \boldsymbol{\delta},$$

where $\boldsymbol{\xi}$, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are independent random vectors of dimensions s, s and t, respectively and $\boldsymbol{\beta}$ is unknown vector of dimension p. Suppose that $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, are n independent observations on this model.

As Gleser pointed out, a naive approach which simply substitutes $\boldsymbol{\xi}$ by $\mathbf{x}$ usually leads to inconsistent estimators of $\boldsymbol{\beta}$. Now let $\bar{\mathbf{x}}$ be a function of $\mathbf{x}$ which will be determined later. We consider the *M*-estimator of $\boldsymbol{\beta}$ by minimizing

$$\sum_{i=1}^n \rho(\mathbf{y}_i - \mathbf{h}(\bar{\mathbf{x}}_i, \boldsymbol{\beta})).$$

In this case, we have $\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = -H(\bar{\mathbf{x}}_i, \boldsymbol{\beta})\, \boldsymbol{\psi}(\mathbf{y}_i - \mathbf{h}(\bar{\mathbf{x}}_i, \boldsymbol{\beta}))$ with $H(\bar{\mathbf{x}}, \boldsymbol{\beta}) = (\partial/\partial\boldsymbol{\beta})\, \mathbf{h}'(\bar{\mathbf{x}}, \boldsymbol{\beta})$. Under the condition on the ρ function made in Section 1, we should have

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = \mathrm{E}H(\bar{\mathbf{x}}_i, \boldsymbol{\beta})[G(\mathbf{h}(\bar{\mathbf{x}}_i, \boldsymbol{\beta}) - \mathbf{h}(\boldsymbol{\xi}_i, \boldsymbol{\beta}_0)) + o(\mathbf{h}(\bar{\mathbf{x}}_i, \boldsymbol{\beta}) - \mathbf{h}(\boldsymbol{\xi}_i, \boldsymbol{\beta}_0))].$$

Thus, if we can choose $\bar{\mathbf{x}}$ such that

$$\mathrm{E}(\mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\beta}_0) \mid \mathbf{x}) = \mathbf{h}(\bar{\mathbf{x}}, \boldsymbol{\beta}_0), \tag{4.7}$$

then, under some minor continuity conditions on H and $\mathbf{h}$, we shall have

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}) = \mathrm{E}(H(\bar{\mathbf{x}}, \boldsymbol{\beta}_0)\, GH'(\bar{\mathbf{x}}, \boldsymbol{\beta}_0))(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

and

$$\mathrm{E}\boldsymbol{\psi}_i(\mathbf{y}_i, \boldsymbol{\beta}_0)\, \boldsymbol{\psi}_i'(\mathbf{y}_i, \boldsymbol{\beta}_0) = \Sigma,$$

where

$$\Sigma = \mathrm{E}(H(\bar{\mathbf{x}}, \boldsymbol{\beta}_0)\, \boldsymbol{\psi}(\boldsymbol{\varepsilon} + \mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\beta}_0) - \mathbf{h}(\bar{\mathbf{x}}, \boldsymbol{\beta}_0))\, \boldsymbol{\psi}'(\boldsymbol{\varepsilon} + \mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\beta}_0) \\ - \mathbf{h}(\bar{\mathbf{x}}, \boldsymbol{\beta}_0))\, H'(\bar{\mathbf{x}}, \boldsymbol{\beta}_0)).$$

Consequently, the conditions of the main theorems are satisfied and hence the main theorems hold true.

Remark. When $\mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\beta})$ is a linear function in both $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$ and the underlying distributions are multivariate normal, the condition (4.7) is

$$\bar{\mathbf{x}} = \mathrm{E}(\boldsymbol{\xi} \mid \mathbf{x}) = \Lambda\mathbf{x} + (I - \Lambda)\, \boldsymbol{\mu}.$$

which is independent of the parameter $\boldsymbol{\beta}$ to be estimated. In case that Λ and $\boldsymbol{\mu}$ are unknown, they can be estimated from the $\mathbf{x}$ observations. Then, these quantities can be replaced by their consistent estimates. We shall not give further details in this remark. We only want to remind the reader that the estimator or $\boldsymbol{\beta}$ defined here are consistent to the true value of the interesting parameter, unlike those defined by Gleser which would converge to something else (see his Theorem 3.1). A key reason to guarantee our approach is applicable is that the function $\bar{\mathbf{x}}$ (see (4.7)) should be independent of $\boldsymbol{\beta}_0$, that is, the function $\bar{\mathbf{x}} = \bar{\mathbf{x}}(\mathbf{x})$ is either completely known, or involves with parameters which is estimable by only the $\mathbf{x}$-sequence. For the least squares approach in linear case, this is possible.

ACKNOWLEDGMENTS

The authors thank the referee for his helpful comments. Y. Wu's research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Bai, Z., Rao, C. R., and Wu, Y. (1992). *M*-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica* **2** 237–254.

[2] Bai, Z., Rao, C. R., and Wu, Y. (1997). *M*-estimation of multivariate linear regression by minimizing the difference of two convex functions. In *Handbook of Statistics*, Vol. 15, to appear.

[3] Basawa, L. V., and Koul, H. L. (1988). Large-sample statistics based on quadratic dispersion. *Internat. Statist. Rev.* **56** 199–219.

[4] Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–433.

[5] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[6] Chhikara, R. S., and Folks, J. L. (1989). *The Inverse Gaussian Distribution*; *Theory, Methodology and Applications*. Dekker, New York.

[7] Gleser, L. J. (1990). Improvements of the Naive approach to estimation in nonlinear Error-in-variables regression models. *Contemporary Mathematics* **112** 99–114.

[8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics—The Approach Based on Influence Functions*. Wiley, New York.

[9] Heiler, S., and Willers, R. (1988). Asymptotic normality of *R*-estimates in the linear model. *Statistics* **19** 173–184.

[10] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[11] Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Univ. Calif. Press, pp. 221–233.

[12] Huber, P. J. (1973). Robust regression. *Ann. Statist.* **1** 799–821.

[13] Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

[14] Martin, R. D. (1980). Robust estimation in autoregressive models. In *Directions in Time Series* Vol. (D. R. Brillinger and G. C. Tiao, Eds.), pp. 228–254. IMS Publication, Haywood CA.

[15] Martin, R. D. (1981). Robust methods for time series. In *Applied Time Series II* (D. F. Findley, Ed.), pp. 683–759. Academic Press, New York.

[16] Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72** 585–592.

[17] Yohai, V. J., and Maronna, R. A. (1979). Asymptotic behavior of *M*-estimators for the linear model. *Ann. Statist.* **7** 258–268.