Statistica Sinica **18**(2008), 405-424

ON CONSTRAINED M-ESTIMATION AND ITS RECURSIVE ANALOG IN MULTIVARIATE LINEAR REGRESSION MODELS

Zhidong Bai[1,2], Xiru Chen[3] and Yuehua Wu[4]

[1]*Northeast Normal University,* [2]*National University of Singapore,*
[3]*Graduate University of the Chinese Academy of Sciences and* [4]*York University*

Abstract: In this paper, the constrained M-estimation of the regression coefficients and scatter parameters in a multivariate linear regression model is considered. Robustness and asymptotic behavior are investigated. Since constrained M-estimation is not easy to compute, an up-dating recursion procedure is proposed to simplify the computation of the estimators when a new observation is obtained. We show that, under mild conditions, the recursion estimates are strongly consistent. A Monte Carlo simulation study of the recursion estimates is also provided.

Key words and phrases: Asymptotic normality, breakdown point, consistency, constrained M-estimation, influence function, linear model, M-estimation, recursion estimation, robust estimation.

1. Introduction

Consider the multivariate linear regression model

$$\boldsymbol{y}_i = B\boldsymbol{x}_i + \boldsymbol{e}_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_i \in \mathcal{R}^m$ are the observed response vectors, $\boldsymbol{x}_i \in \mathcal{R}^p$ are the covariate vectors, and $\boldsymbol{e}_i \in \mathcal{R}^m$ are identically and independently distributed (i.i.d.) random error vectors with zero mean vector and covariance matrix V. We focus on the robust estimation of B and V in this paper.

It is noted that when $B = \boldsymbol{\mu} \in \mathcal{R}^m$ and $\boldsymbol{x}_i = 1$, (1) becomes the multivariate location model; when $m = 1$ and $B^T = \boldsymbol{\beta} \in \mathcal{R}^p$, (1) is a usual linear regression model.

There is a rich literature on estimation of the regression coefficients and scatter parameters for the model (1). The well-known method is Least Squares. Even though this method is efficient for normal distributed errors and is mathematically convenient, it is not resistant to outliers nor stable with respect to deviations from the given distributional model. Many robust statistical procedures have then been developed.

Two standard classes of robust estimates are M-estimates and S-estimates (Kent and Tyler (1996)), M-estimators can be tuned to have such local robustness properties, as efficiency and a bound on the influence function at an underlying distribution such as the multivariate normal. However, M-estimates suffer from poor breakdown properties in high dimensions. On the other hand, S-estimates can be tuned to have good breakdown properties, but when tuned in this way, they tend to suffer from poor local robustness. Note that the robustness of M-estimators of regression coefficients, and especially of location parameters, is determined by the increasing rate of the dispersion function. Thus to pursue most robust estimators, bounded dispersion functions are popularly employed. However, when a dispersion function is bounded, the minimization problem would have no solution if affine equivariant estimation is considered, because the objective function will tend to negative infinity when the smallest eigenvalue of the estimated scatter matrix tends to 0. Hence Kent and Tyler (1996) proposed a hybrid estimation, called constrained M-estimation (CM-estimation), for estimating the location and scatter parameters in a multivariate location model, and which achieves both good local and global robustness besides being affine equivariant. Later, Mendes and Tyler (1996) examined CM-estimates in a linear regression model, that are regression equivariant: affine equivariant to $\boldsymbol{x}$ and affine equivariant to $\boldsymbol{y}$. In this paper, we will further consider CM-estimation for B and V in the model (1).

Let $Z_n = \{(\boldsymbol{y}_1^T, \boldsymbol{x}_1^T), \ldots, (\boldsymbol{y}_n^T, \boldsymbol{x}_n^T)\}$ be a data set in $\mathcal{R}^{m+p}$ and let $\mathcal{P}_m$ denote the set of all $m \times m$ positive definite symmetric matrices. For the data set Z_n and a prechosen constant $\varepsilon \in (0,1)$, the CM-estimates of B and V, denoted by $\hat{B}(Z_n) \in \mathcal{R}^{m\times p}$ and $\hat{V}(Z_n) \in \mathcal{P}_m$, are any pairs that minimize the objective function

$$L(B,V;Z_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\rho\Big\{(\boldsymbol{y}_i - B\boldsymbol{x}_i)^T V^{-1}(\boldsymbol{y}_i - B\boldsymbol{x}_i)\Big\} + \frac{1}{2}\log\{\det(V)\}\right] \quad (2)$$

over all $B \in \mathcal{R}^{m\times p}$ and $V \in \mathcal{P}_m$, subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\left[\rho\Big\{(\boldsymbol{y}_i - B\boldsymbol{x}_i)^T V^{-1}(\boldsymbol{y}_i - B\boldsymbol{x}_i)\Big\}\right] \le \varepsilon\rho(\infty), \quad (3)$$

where $\rho(s)$ is a bounded nondecreasing function for $s \ge 0$. In general, the minimization problem which defines the CM-estimates may have multiple solutions. We use the notation $(\hat{B}(Z_n), \hat{V}(Z_n))$ to refer a measurable solution to the minimization problem rather than to the set of all solutions.

It can be shown that the CM-estimates defined above are regression equivariant.

The properties of CM-estimation are to be investigated. It is noted that most M-estimates have no explicit expressions. Often the Newton approach cannot be applied to the computation of the parameter estimates, but, even when it can, it is usually sensitive to initial values. Moreover, when a new observation is obtained, it is not easy to recalculate the estimates. There is often a need to update CM-estimates when new observations are obtained. The first attempt in this direction was made in Bickel (1975) with the so called "one-step approximation". Among other such attempts, one is recursive estimation of the parameters based on the previous estimates and a new observation. Recursive estimates can be easy to calculate and do not require extensive storage of data. In the case of M-estimation, see Englund, Holst and Ruppert (1988), Englund (1993), Bai and Wu (1993), Miao and Wu (1996) and Wu (1996), among others. Note that such up-dating is also important to on-line learning in neural computation. The computation of CM-estimates $\hat{B}$ and $\hat{V}$ is even more complicated; it is presented here and we study its asymptotic behavior.

The organization of this paper is as follows: In Section 2, we study the existence of the CM-estimates and CM-functionals. In Section 3, we consider the finite sample breakdown point of the CM-estimates. In Section 4, the consistency, influence functions and asymptotic normality of the CM-estimates are investigated. In Section 5, a recursive computation of the CM-estimates is proposed. The recursive estimates are shown to be strongly consistent under mild conditions. Some simulation results are presented in Section 6. Proofs of the theorems in this paper are given in the Appendix.

2. Existence of the CM-estimates and CM-functionals

Throughout this paper we assume that ρ satisfies the following assumption.

Assumption 2.1. For $t \geq 0$, ρ is nondecreasing, $0 = \rho(0) < \rho(\infty) < \infty$ and $\rho(t)$ is continuous from above at 0.

As in Kent and Tyler (1996), results on existence can be made more general by introducing the notion of CM-functionals.

Assume that the random vector $(\boldsymbol{y}^T, \boldsymbol{x}^T) \in \mathcal{R}^{m+p}$ has the joint distribution F. With $0 < \varepsilon < 1$ being fixed, we define the CM-functionals for $B(F)$ and $V(F)$ in a manner analogous to (2) and (3), i.e., $(B(F), V(F))$ minimizes over all $(B, V) \in \mathcal{R}^{m\times p} \times \mathcal{P}_m$ the objective function

$$\mathcal{L}(B, V) = E\Big[\rho\{(\boldsymbol{y} - B\boldsymbol{x})^T V^{-1}(\boldsymbol{y} - B\boldsymbol{x})\}\Big] + \frac{1}{2}\log\{\det(V)\} \tag{4}$$

subject to the constraint

$$E\Big[\rho\{(\boldsymbol{y} - B\boldsymbol{x})^T V^{-1}(\boldsymbol{y} - B\boldsymbol{x})\}\Big] \leq \varepsilon\rho(\infty). \tag{5}$$

For establishing the existence of the CM-functionals, we make the following assumption on F, the joint distribution of $(\boldsymbol{y}^T, \boldsymbol{x}^T)$.

Assumption 2.2.
(a) For any $B \in \mathcal{R}^{m\times p}$ and any hyperplane H in $\mathcal{R}^m$, $\Pr\{(\boldsymbol{y} - B\boldsymbol{x}) \in H\} < 1-\varepsilon$.
(b) For any $B \in \mathcal{R}^{m\times p}$, $B \neq 0$, $\Pr\{B\boldsymbol{x} = \boldsymbol{0}\} < 1 - \varepsilon$.
(c) For any $B \in \mathcal{R}^{m\times p}$, $\Pr\{\boldsymbol{y} = B\boldsymbol{x}\} < 1 - \varepsilon$.

Remark 2.1. It can be seen that Assumption 2.2(c) is a consequence of Assumption 2.2(a).

Remark 2.2. One can show that Assumption 2.2(a) implies the existence of a positive constant δ such that, for any $B \in \mathcal{R}^{m\times p}$ and any hyperplane H in $\mathcal{R}^m$, $\Pr\{(\boldsymbol{y} - B\boldsymbol{x}) \in H\} < 1 - \varepsilon - \delta$. Similarly, under Assumption 2.2(b), there exists a positive δ such that $1 - \varepsilon$ can be replaced by $1 - \varepsilon - \delta$ in Assumption 2.2(b).

We have the following theorem on the existence of CM-functionals. Remark 2.2 is needed in its proof.

Theorem 2.1. *Under Assumptions* 2.1, 2.2(*a*) *and* (*b*), *there exists a* $(B_0, V_0) \in \mathcal{R}^{m\times p} \times \mathcal{P}_m$ *that minimizes* $\mathcal{L}(B, V)$ *subject to the constraint* (5). *Furthermore, the set of all such* (B_0, V_0) *is bounded away from* $\partial(\mathcal{R}^{m\times p} \times \mathcal{P}_m)$, *the boundary set of* $\mathcal{R}^{m\times p} \times \mathcal{P}_m$.

The following theorem shows that Assumption 2.2(c) is a necessary condition of the existence of the CM-functionals, and it also plays an important role in the next section in obtaining the breakdown point of the CM-estimates. Its proof is given in the Appendix.

Theorem 2.2. *Under Assumption* 2.1, *if Assumption* 2.2(*c*) *does not hold, then there does not exist* $(B_0, V_0) \in \mathcal{R}^{m\times p} \times \mathcal{P}_m$ *which minimizes* $\mathcal{L}(B, V)$ *subject to the constraint* (5).

Let F_n denote the empirical distribution of $\{(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. Then the CM-functionals at F_n correspond to the CM-estimates defined by (2) under the constraint (3) for the data set.

We use the following notation:

$$C_{1,n}^{(m,p)} = \max_{B\in\mathcal{R}^{m\times p},\ H\subset\mathcal{R}^m} \#\{i : (\boldsymbol{y}_i - B\boldsymbol{x}_i) \in H, 1 \le i \le n\}, \tag{6}$$

$$C_{2,n}^{(m,p)} = \max_{B\in\mathcal{R}^{m\times p},\ B\neq 0} \#\{i : B\boldsymbol{x}_i = \boldsymbol{0}, 1 \le i \le n\}, \tag{7}$$

$$C_{3,n}^{(m,p)} = \max_{B\in\mathcal{R}^{m\times p}} \#\{i : \boldsymbol{y}_i = B\boldsymbol{x}_i, 1 \le i \le n\}, \tag{8}$$

where H is an arbitrary hyperplane in $\mathcal{R}^m$. Note that $C_{1,n}^{(m,p)}$, $C_{2,n}^{(m,p)}$, and $C_{3,n}^{(m,p)}$ are, respectively, n times the maximum values of $\Pr\{(\boldsymbol{y} - B\boldsymbol{x}) \in H\}$, $\Pr\{B\boldsymbol{x} =$

$\mathbf{0}\}$, and $\Pr\{\boldsymbol{y} = B\boldsymbol{x}\}$ for the empirical distribution F_n, defined in Assumption 2.2. Therefore Assumption 2.2(a), (b), and (c) hold for F_n if $C_{1,n}^{(m,p)} < n(1-\varepsilon)$, $C_{2,n}^{(m,p)} < n(1-\varepsilon)$, and $C_{3,n}^{(m,p)} < n(1-\varepsilon)$. By the Strong Law of Large Numbers, when Assumption 2.2 holds, the above conditions hold for all large n with probability 1. Accordingly the results of the above theorems apply to CM-estimates. Note that by Remark 2.1, $C_{3,n}^{(m,p)} \leq C_{1,n}^{(m,p)}$.

3. Finite Sample Breakdown Point

In this section, we study the breakdown point of the CM-estimates using the finite replacement breakdown point introduced by Donoho and Huber (1983). For a fixed n, suppose that $n_1 (\leq n)$ is the smallest integer such that a replacement of n_1 data points from the original data set $Z_n = \{(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)\}$ will lead to a corrupted sample $\tilde{Z}_n$. That is, for a given $\varepsilon_{n_1} = n_1/n$, the statistic $(\hat{B}(\cdot), \hat{V}(\cdot))$ is said to break down at Z_n under ε_{n_1}- contamination if replacing n_1 data points by $Z_{n_1}^*$ allows one of the following to happen:

(i)$(\hat{B}(\tilde{Z}_n), \hat{V}(\tilde{Z}_n))$ does not exist for a choice of $Z_{n_1}^*$;
(ii)The supreme of $\|\hat{B}(\tilde{Z}_n)\|$ for possible choices of $Z_{n_1}^*$ is infinity;
(iii)The supremum of the largest eigenvalue of $\hat{V}(\tilde{Z}_n)$ for possible choices of $Z_{n_1}^*$ is infinity, or the infimum of smallest eigenvalue of $\hat{V}(\tilde{Z}_n)$ for possible choices of $Z_{n_1}^*$ is 0.

The finite sample replacement breakdown point of $(\hat{B}(\cdot), \hat{V}(\cdot))$ at Z_n is then defined to be $\varepsilon_n^*(Z_n)$, the minimum of all ε_{n_1} causing breakdown.

The bounds on the number of the finite sample replacement breakdown points are given by the following theorem. Its proof is provided in the Appendix.

Theorem 3.1. *If ρ satisfies Assumption* 2.1, $0 < \varepsilon < 1$, *and* $C_n = \max(C_{1,n}^{(m,p)}, C_{2,n}^{(m,p)}) < n(1-\varepsilon)$ *for the data set Z_n, then we have*

$$\min\left\{\frac{\lceil n\varepsilon\rceil}{n}, \frac{\lceil n(1-\varepsilon)\rceil - C_n}{n}\right\} \leq \varepsilon_n^*(Z_n) \leq \min\left\{\frac{\lfloor n\varepsilon\rfloor + 1}{n}, \frac{\lceil n(1-\varepsilon) - C_{3,n}^{(m.p)}\rceil}{n}\right\},$$

where $C_{1,n}^{(m,p)}, C_{2,n}^{(m,p)}$ and $C_{3,n}^{(m.p)}$ are defined as before, $\lceil\kappa\rceil$ represents the smallest integer greater than or equal to κ if positive and zero otherwise, and $\lfloor\kappa\rfloor$ represents the largest integer less than or equal to κ if positive, and is zero otherwise.

4. Consistency, Influence Functions and Asymptotic Normality

The following discussion is based on the results of Kent and Tyler (1996).

We assume throughout this section that the $(m+p)$-dimensional random vector $(\boldsymbol{y}^T, \boldsymbol{x}^T)$ has an absolutely continuous distribution F in $\mathcal{R}^{m+p}$. We also assume that the following uniqueness of the CM-functional condition holds.

Assumption 4.1. The CM-functional $(B(F), V(F))$ exists and is uniquely defined at F.

Remark 4.1. Note that $(B(F), V(F))$ is shown to exist under Assumptions 2.1−2.2 in Section 2.

Assumption 4.2. Assumption 2.1 holds and ρ is continuous for $t \geq 0$.

Suppose that Assumptions 4.1 and 4.2 hold. If $F_k \xrightarrow{w} F$, then it follows that

$$(B(F_k), V(F_k)) \to (B(F), V(F)).$$

This result also warrants the strong consistency of the CM-estimates since the empirical distribution converges to the underlying distribution function almost surely.

Given weak consistency, the influence functions and the asymptotic distributions of the CM-estimates derive from the following estimating equations. Let $s = (\boldsymbol{y} - B\boldsymbol{x})^T V^{-1} (\boldsymbol{y} - B\boldsymbol{x})$. Assume that ρ is differentiable. Then $(B(F), V(F))$ must satisfy the estimating equations:

$$E[\rho'(s)(\boldsymbol{y} - B\boldsymbol{x})\boldsymbol{x}^T] = 0 \tag{9}$$

$$V = \frac{mE[\rho'(s)(\boldsymbol{y} - B\boldsymbol{x})(\boldsymbol{y} - B\boldsymbol{x})^T]}{E[s\rho'(s)]}, \tag{10}$$

with either

$$E[2s\rho'(s)] = m \tag{11}$$

or

$$E[\rho(s)] = \varepsilon\rho(\infty). \tag{12}$$

The empirical versions of (9)−(12) are given by

$$\sum_{i=1}^{n} \rho'(s_i)(\boldsymbol{y}_i - B\boldsymbol{x}_i)\boldsymbol{x}_i^T = 0, \tag{13}$$

$$V = \frac{m \sum_{i=1}^{n} \left[\rho'(s_i)(\boldsymbol{y}_i - B\boldsymbol{x}_i)(\boldsymbol{y}_i - B\boldsymbol{x}_i)^T\right]}{\sum_{i=1}^{n} \{s_i \rho'(s_i)\}}, \tag{14}$$

$$\frac{1}{n} \sum_{i=1}^{n} \{2s_i \rho'(s_i)\} = m, \tag{15}$$

$$\frac{1}{n} \sum_{i=1}^{n} \{\rho(s_i)\} = \varepsilon\rho(\infty), \tag{16}$$

where $s_i = (\boldsymbol{y}_i - B\boldsymbol{x}_i)^T V^{-1}(\boldsymbol{y}_i - B\boldsymbol{x}_i)$.

It can be seen that Equations (13), (14) and (15) hold whenever strict inequality holds in (3) for the CM-estimates, and they arise as the critical points of (2). In contrast, Equations (13), (14) and (16) hold whenever equality holds in (3) for the CM-estimates, and they arise as the critical points of (2) after introducing a Lagrange multiplier to account for the constraint (16).

We note that (B, V) consists of $l = mp + (1/2)m(m+1)$ parameters. Let $\boldsymbol{\theta} \in \mathcal{R}^l$ represent an l-dimensional vector parameterization of (B, V) with B being the first mp components of $\boldsymbol{\theta}$ and with the upper triangular part of V being the remaining $(1/2)m(m+1)$ components. Analogously, the CM-functionals $(B(F), V(F))$ can be represented by $\boldsymbol{\theta}(F) \in \mathcal{R}^l$. It is easy to see that there exists a function $\Psi_1 : \mathcal{R}^{m+p} \times \mathcal{R}^l \to \mathcal{R}^l$ such that (9), (10) and (11) together are equivalent to

$$E[\Psi_1(F, \boldsymbol{\theta})] = \mathbf{0}, \tag{17}$$

and there exists a function $\Psi_2 : \mathcal{R}^{m+p} \times \mathcal{R}^l \to \mathcal{R}^l$ such that (9),(10) and (12) together are equivalent to

$$E[\Psi_2(F, \boldsymbol{\theta})] = \mathbf{0} \tag{18}$$

with the first $(l-1)$ entries of Ψ_1 and Ψ_2 being the same.

Now we can obtain the local properties of the CM-estimates under the following assumptions.

Assumption 4.3. The function ρ has a continuous second derivative, and $\rho'(t)$, $t\rho'(t)$, and $\rho''(t)$ are bounded.

Assumption 4.4. For the CM-functional $(B(F), V(F))$, (11) and (12) do not both hold at F.

Note that under Assumptions 4.1, 4.2 and 4.4, if $F_k \to F$ in distribution, then (11) and (12) cannot both hold for $(B(F_k), V(F_k))$ for all large k. Furthermore, for all large k, if (11) holds for F, then (11) holds for F_k and if (12) holds for F, then (12) holds for F_k. As shown in Kent and Tyler (1996), this allows one to treat (17) and (18) as two cases when studying the local properties of the CM-estimates. For convenience, we treat both cases together by defining $\Psi = \Psi_1$ if (11) holds and $\Psi = \Psi_2$ if (12) holds.

The influence function of the functional $\boldsymbol{\theta}(F)$, the $\mathcal{R}^l$ representation of $(B(F), V(F))$, at $(\boldsymbol{y}_0, \boldsymbol{x}_0)$, is defined by

$$IF(\boldsymbol{y}_0, \boldsymbol{x}_0; \boldsymbol{\theta}(F)) = \lim_{h \to 0^+} \frac{\boldsymbol{\theta}(F_h(\boldsymbol{y}_0, \boldsymbol{x}_0)) - \boldsymbol{\theta}(F)}{h}$$

provided that the limit exists, where $F_h(\boldsymbol{y}_0, \boldsymbol{x}_0) = (1-h)F + h\delta_{(\boldsymbol{y}_0, \boldsymbol{x}_0)}$, with $\delta_{(\boldsymbol{y}_0, \boldsymbol{x}_0)}$ denoting the atomic probability distribution concentrated at $(\boldsymbol{y}_0, \boldsymbol{x}_0) \in \mathcal{R}^{m+p}$.

Let $\lambda(\boldsymbol{\theta}) = E[\Psi(F, \boldsymbol{\theta})]$. Suppose that $\lambda(\boldsymbol{\theta})$ has a nonsingular derivative $\Lambda = \partial\lambda(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta}(F)$. Under Assumptions 4.1, 4.2, 4.3 and 4.4, the influence function of $\boldsymbol{\theta}(F)$ can be shown to exist and is given by $IF(\boldsymbol{y}_0, \boldsymbol{x}_0; \boldsymbol{\theta}(F)) = -\Lambda^{-1}\Psi(F, \boldsymbol{\theta}(F))$. The limiting distribution of $\hat{\boldsymbol{\theta}}(Z_n)$ is given next.

Theorem 4.1. *Under Assumptions* 4.1–4.4, *if* $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_1)$ *are i.i.d.* $(\boldsymbol{y}, \boldsymbol{x})$, *then*

$$\sqrt{n}\{\hat{\boldsymbol{\theta}}(Z_n) - \boldsymbol{\theta}(F)\} \to \mathbf{N}_l(\mathbf{0}, \Lambda^{-1}M(\Lambda^T)^{-1}) \quad \textit{in distribution},$$

where M *is the variance-covariance matrix of* $\Psi(F, \boldsymbol{\theta}(F))$, *and* Λ *is defined above.*

Proof. The proof mimics the proof of Theorem 4.1 given by Lopuhaä (1989).

5. Recursive Estimation

As indicated in Section 1, CM-estimation is difficult to implement and other approaches need to be explored. Motivated by Rubinstein (1986) Englund (1993), Bai and Wu (1993) and Miao and Wu (1996), we propose the following up-dating recursion estimation of B and V.

$$\begin{cases} \tau_{n+1} = \max\left\{0, \tau_n + \nu_n\Big(\dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n}\rho(\|\boldsymbol{y}_{i+1} - B_i\boldsymbol{x}_{i+1}\|^2_{\tilde{V}_i}) - \varepsilon\rho(\infty)\Big)\right\}, \\ B_{n+1} = B_n + a_n(1+\tau_{n+1})H_1(B_n, V_n, \boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1})\tilde{S}^{-1}_{n+1}, \\ V_{n+1} = V_n + (n+1)^{-1}(1+\tau_{n+1})H_2(B_n, V_n, \boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}), \end{cases} \tag{19}$$

where $h(t) = (d\rho(t^2)/dt)/(t_0^2 t)$, $\tilde{S}_n = \sum_{i=1}^{n} v_i h(\|\boldsymbol{y}_i - B_i\boldsymbol{x}_i\|_{\tilde{V}_i})\boldsymbol{x}_i\boldsymbol{x}_i^T$, and

$$H_1(B, V, \boldsymbol{x}, \boldsymbol{y}) = h(\|\boldsymbol{y} - B\boldsymbol{x}\|_{\tilde{V}})(\boldsymbol{y} - B\boldsymbol{x})\boldsymbol{x}^T,$$
$$H_2(B, V, \boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{y} - B\boldsymbol{x})(\boldsymbol{y} - B\boldsymbol{x})^T\frac{h(\|\boldsymbol{y} - B\boldsymbol{x}\|_{\tilde{V}})}{2} - V,$$

B_0 is arbitrary, V_0 is a positive definite matrix, $\{a_n\}$ satisfies certain conditions, and $\tilde{V}$ is a Lipschitz continuous $m \times m$ matrix function of V defined as follows:

> Let λ_i and $\boldsymbol{\alpha}_i$ be the i-th eigenvalue and corresponding eigenvector of V respectively. Then
>
> $$\tilde{V} = \sum_{i=1}^{m}\tilde{\lambda}_i\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T,$$
>
> where $\tilde{\lambda}_i = (\delta_1 \vee \lambda_i) \wedge \delta_2$ and δ_1, δ_2 $(0 < \delta_1 < \delta_2 < \infty)$ are two appropriate constants.

Here t_0 is defined in Assumption 5.4, which will be given later. It is assumed that $v_n = g(\lfloor 0.00001n \rfloor) > 0$, $n = 1, 2, \ldots$, are chosen such that $\tilde{S}_n/n \to A > 0$. Note that v_n in (19) is a constant sequence satisfying certain conditions.

When the sample size is not too small, this estimation is shown to perform very well in the simulation study in Section 6.

Note that (19) is a recursive analog of the unconstrained M-estimation, which minimizes (2), if $\tau_n \equiv 0$ in (19).

We make the following assumptions to facilitate the investigation of the limiting behavior of the recursion estimators given by (19).

Assumption 5.1. $(\boldsymbol{x}_i, \boldsymbol{e}_i)$, $i = 1, 2, \ldots$ are i.i.d. copies of $(\boldsymbol{x}, \boldsymbol{e})$, and $\boldsymbol{x}$ and $\boldsymbol{e}$ are independent. The distribution of $\boldsymbol{e}_i$ has a density $|\Sigma|^{-1/2} f_1(\|\boldsymbol{e}\|_\Sigma)$, where f_1 is decreasing on $[0, \infty)$ and strictly decreasing in a neighborhood of 0. The random vector $\boldsymbol{x}_i$ has finite second moment denoted by $G = E\boldsymbol{x}_i\boldsymbol{x}_i^T > 0$. In the sequel, F_2 stands for the common distribution of $\boldsymbol{x}_i$.

Let $\{n(k)\}$ be a sequence of increasing positive integers satisfying

$$\sum_{k=1}^{\infty} (n(k)d_k)^{-1} < \infty, \text{ and } d_k \to 0 \text{ as } k \to \infty, \tag{20}$$

where $d_k = \sum_{\ell=n(k)+1}^{n(k+1)} \ell^{-1}$. Then we have

$$\sum_{k=1}^{\infty} d_k = \infty. \tag{21}$$

As an example, one can choose $n(k) = [k^\delta]$ for some $\delta > 2$, where $[c]$ denotes the integer part of c.

Assumption 5.2. The sequence $\{a_n, n = 1, 2, \ldots\}$ is adaptive, i.e., a_n is $\mathcal{F}_n$-measurable, $\mathcal{F}_n$ the sigma field generated by the random vectors $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)^T$, $i = 1, \ldots, n$, and $0 < \pi_0 \le a_n \le \pi_1 < \infty$ for some π_0 and π_1 and all n.

Assumption 5.3. $0 < \nu_n \le O(n^{-\upsilon})$ and $\upsilon > 1$.

Assumption 5.4. There exists $t_0 > 0$ such that h is constant for $t > t_0$ and $h(t) < h(t_0)$ for $t < t_0$. $th(t)$ is uniformly continuous for $t < t_0$. $h(0) \neq 0$.

For describing Assumption 5.5, we need the following notation:

Suppose that f_1 and u are given as above. Let $\omega > 0$ be the solution of the equation

$$2m = \int \omega \boldsymbol{z}^T \boldsymbol{z} h(\omega \|\boldsymbol{z}\|^2) f_1(\|\boldsymbol{z}\|^2) d\boldsymbol{z},$$

where m is the dimension of dependent variable $\boldsymbol{z}$. Let $\Omega = \omega^{-1}\Sigma$. By Bai and Wu (1993), Ω satisfies

$$\Omega = E\boldsymbol{e}\boldsymbol{e}^T \frac{h(\|\boldsymbol{e}\|_\Omega^2)}{2}. \tag{22}$$

Define

$$\zeta = \left\{ \frac{1}{8m} \int \left[\omega \|\boldsymbol{z}\|^2 h(\omega \|\boldsymbol{z}\|^2) - \frac{\omega}{1.5} \|\boldsymbol{z}\|^2 h\Big(\frac{\omega}{1.5} \|\boldsymbol{z}\|^2 \Big) \right] f_1(\|\boldsymbol{z}\|^2) d\boldsymbol{z} \right\} \wedge 0.1.$$

Assumption 5.5. Suppose that $\delta_1 < \zeta \lambda_*(\Omega)$ and $\delta_2 > 3\lambda^*(\Omega)$, where $\lambda_*(\Omega)$ and $\lambda^*(\Omega)$ are the smallest and largest eigenvalue of Ω, δ_1 and δ_2 are used to define $\widetilde{V}_n$ in (19), and ζ is given above.

The following theorem states that (B_n, V_n) is strongly consistent. Its proof is provided in the Appendix.

Theorem 5.1. *Under Assumptions* 5.1–5.4, *B_n is a strongly consistent estimate of B. Further, with Assumption* 5.5, *V_n is a strongly consistent estimate of Ω.*

It is noted that f_1 is usually unknown and so is Ω. How to obtain a good estimator of Σ based on (19) is an interesting problem. In order to estimate Σ, we propose to add the following to (19) for $n \geq n_0$:

$$\begin{cases} S_{n+1} = S_n + (\boldsymbol{y}_{n+1} - B_n \boldsymbol{x}_{i+1})(\boldsymbol{y}_{n+1} - B_n \boldsymbol{x}_{i+1})^T I\Big[\rho(\|\boldsymbol{y}_{n+1} - B_n \boldsymbol{x}_{n+1}\|^2_{\widetilde{V}_n}) \neq 0 \Big], \\ k_{n+1} = k_n + I\Big[\rho(\|\boldsymbol{y}_{n+1} - B_n \boldsymbol{x}_{n+1}\|^2_{\widetilde{V}_n}) \neq 0 \Big], \\ \Sigma_{n+1} = \dfrac{S_{n+1}}{k_{n+1}}, \end{cases} \tag{23}$$

with $k_{n_0} = 1$, where $I(A)$ is an indicator function on a set A. Since B_n is strongly consistent, it can be shown that Σ_n is a consistent estimator of Σ.

6. Simulation Study

In this section, we study the finite sample performance of the recursion estimation given by (19). Let ρ be Tukey's biweight function

$$\rho(\tilde{t}^2) = c \times \begin{cases} \dfrac{\tilde{t}^2}{2} - \dfrac{\tilde{t}^4}{2} + \dfrac{\tilde{t}^6}{6}, & \text{for } \ 0 < \tilde{t} \leq 1, \\ \dfrac{1}{6}, & \text{for } \ \tilde{t} > 1, \end{cases}$$

where $\tilde{t} = t/t_0$.

Hence,

$$h(\tilde{t}) = c \times \begin{cases} 1 - 2\tilde{t}^2 + \tilde{t}^4, & \text{for } \ 0 < \tilde{t} \leq 1, \\ 0, & \text{for } \ \tilde{t} > 1. \end{cases}$$

In the following simulation, $c = 2$, $t_0 = 100$, $\nu_n = 1/n^2$, $a_n = 0.9$, $v_n = 1$, $\delta_1 = 0.1$, $\delta_2 = 5$, $\varepsilon = 0.5$ and $B = \begin{pmatrix} 5 & -6 \\ 0 & 7 \\ -9 & 3 \end{pmatrix}$.

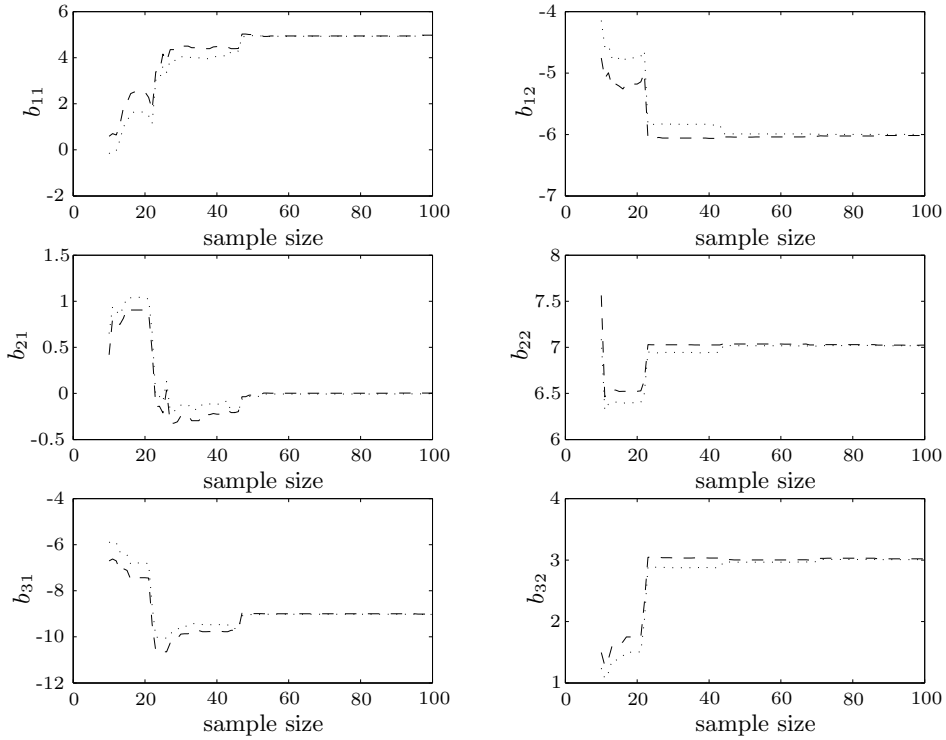

Figure 1. The estimates of regression coefficients: - - - - - by (19); · · · · · by (19) with $\tau_i = 0$ for all i (unconstrained case) for Tukey's biweight function with $c = 2$.

Let $\boldsymbol{x}_i$, $i = 1, 2, \ldots$, be generated as follows: First generate $\boldsymbol{x}_i$ from the distribution $N\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$. Then select an element $\boldsymbol{x}_i$ randomly with equal probability and add a value generated from $N(100, 10)$ to it with probability 0.8. Let

$$A = \begin{pmatrix}0.8 & 0 & 0\\0.2 & 1.2 & 0\\0.1 & 0.4 & 1.1\end{pmatrix}$$

and $\Sigma^{-1} = AA^T$. The random error vectors $\boldsymbol{e}_i$, $i = 1, 2, \ldots$, are generated as follows. First generate $\boldsymbol{e}_i$ from the distribution $0.6N(\mathbf{0}, \Sigma) + 0.4N(\mathbf{0}, 36\Sigma)$. Then select an element of $\boldsymbol{e}_i$ randomly with equal probability and add a value generated from $N(1,000, 100)$ to it with probability 0.8. We calculate $\boldsymbol{y}_i$, $i = 1, 2, \ldots$, as

$$\boldsymbol{y}_i = B\boldsymbol{x}_i + \boldsymbol{e}_i, \quad i = 1, 2, \ldots.$$

The simulation results are reported in Figures 1−4. From Figures 1−2, it can be seen that for Tukey's biweight function, CM-estimates seem to perform

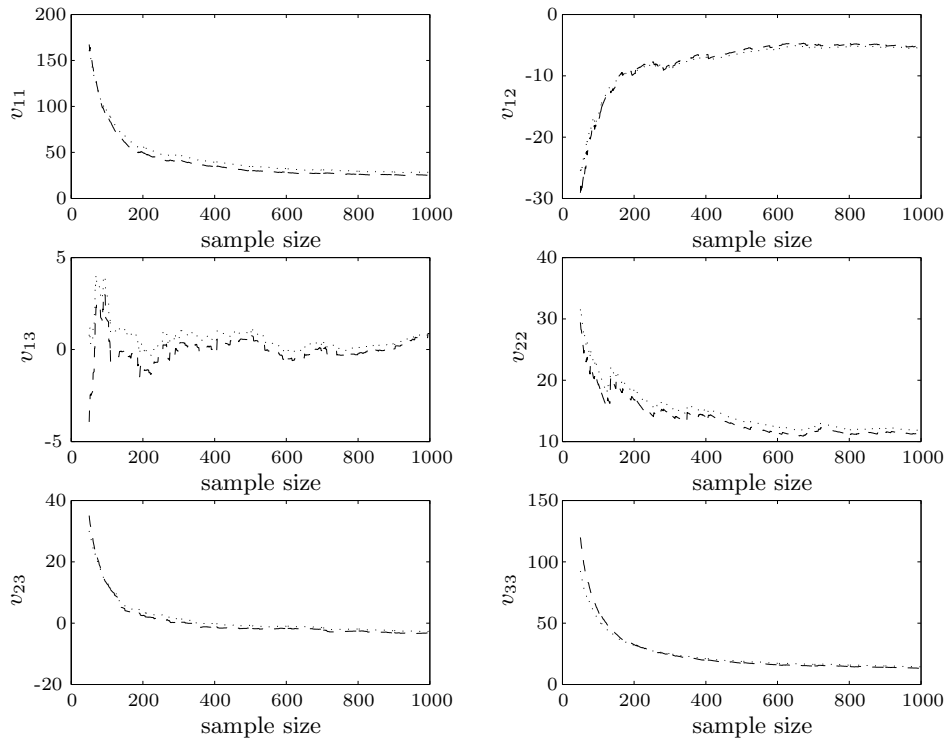

Figure 2. The estimates of scatter parameters: - - - - - by (19); · · · · · by (19) with $\tau_i = 0$ for all i (unconstrained case) for Tukey's biweight function with $c = 2$.

better than the M-estimates. Figure 3 shows that the estimates of σ_{ij} given by (23) tend to the true values when the sample size increases. From Figure 4, it can be observed that the recursion estimation (19) using Tukey's biweight function with $c = 2$ seems to perform much better than the recursion estimation (19) with $\tau_i = 0$ using Huber ρ function with $c = 1.345$ for estimating the regression coefficients.

Appendix

Proof of Theorem 2.1. It is clear that if $(B, V) \in \mathcal{R}^{m\times p} \times \mathcal{P}_m$, then $\mathcal{L}(B, V) < \infty$. For all large λ, $(B, \lambda V) \in \mathcal{R}^{m\times p} \times \mathcal{P}_m$ satisfies (5).

Therefore, to complete the proof it suffices to show that if a sequence of $(B_k, V_k) \to \partial(\mathcal{R}^{m\times p} \times \mathcal{P}_m)$, then either $\mathcal{L}(B_k, V_k) \to \infty$ or the constraint (5) is not met for large k.

Let $\lambda_{1,k}$ and $\lambda_{m,k}$ represent the largest and smallest eigenvalues of V_k respectively. If constraint (5) holds for some (B_k, V_k), then for any $\tilde{s} \geq 0$ and

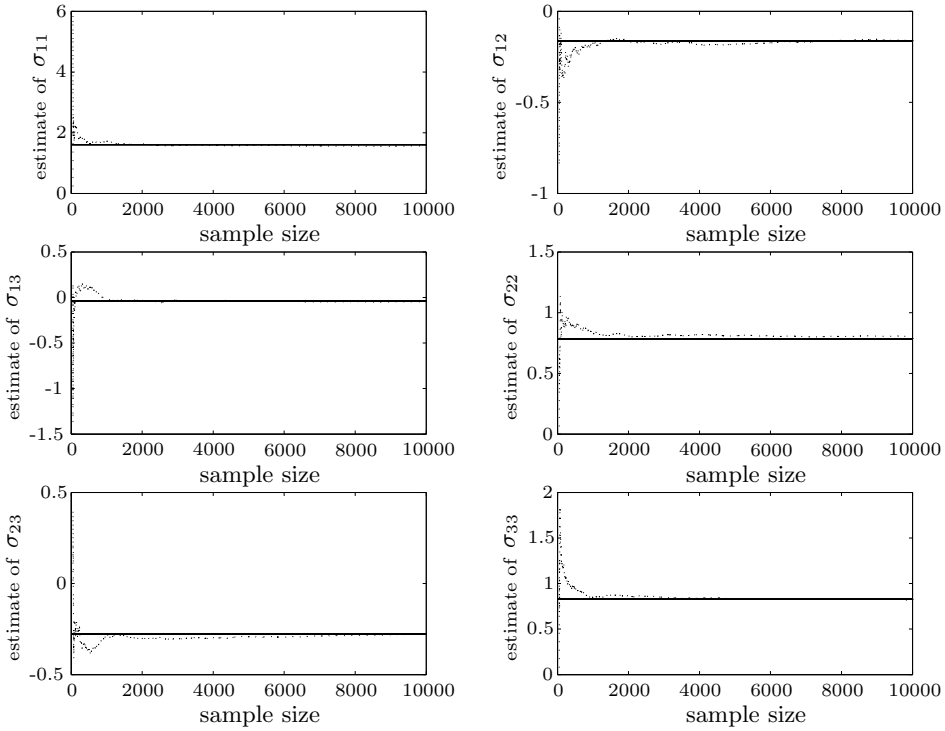

Figure 3. ———— true values of σ_{ij}; · · · · · estimates of σ_{ij} by (23) for Tukey's biweight function with $c = 2$.

$$A_k = \{(\boldsymbol{x}, \boldsymbol{y}) : (\boldsymbol{y} - B_k\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B_k\boldsymbol{x}) \geq \tilde{s}\},$$

$$\varepsilon\rho(\infty) \geq \int_{A_k} \rho\{(\boldsymbol{y} - B_k\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B_k\boldsymbol{x})\} dF(\boldsymbol{y}, \boldsymbol{x}) \geq \rho(\tilde{s}) \Pr(A_k).$$

Thus, for any $\delta > 0$, we can choose $\tilde{s}_0$ being so large that

$$\rho(\tilde{s}_0) \geq \frac{\varepsilon}{\varepsilon + \delta}\rho(\infty),$$

and hence

$$\Pr\{(\boldsymbol{y} - B_k\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B_k\boldsymbol{x}) < \tilde{s}_0\} \geq 1 - \varepsilon - \delta. \tag{24}$$

This in turn implies that

$$\Pr\left\{[\boldsymbol{a}_k^T(\boldsymbol{y} - B_k\boldsymbol{x})]^2 < \lambda_{m,k}\tilde{s}_0\right\} \geq 1 - \varepsilon - \delta, \tag{25}$$

where $\boldsymbol{a}_k$ is an eigenvector of V_k associated with $\lambda_{m,k}$ and is normalized so that $\boldsymbol{a}_k^T\boldsymbol{a}_k = 1$.

If $\lambda_{m,k} \to 0$ along some subsequence, then (25) will lead to a contradiction to Remark 2.2. Thus, we may assume $\lambda_{m,k} > \lambda_m > 0$.

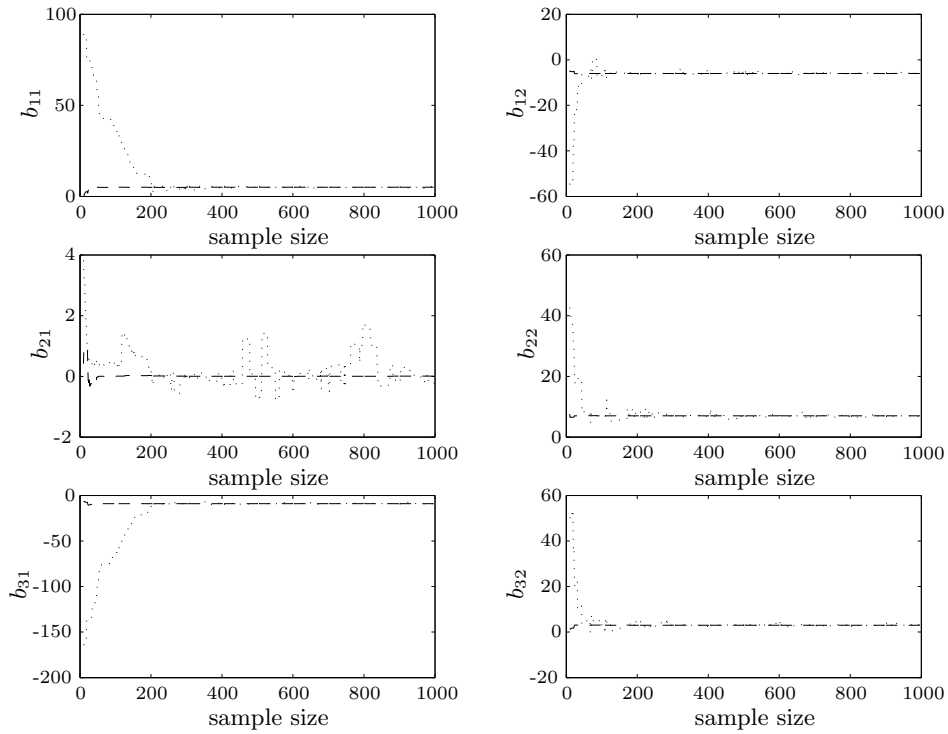

Figure 4. The estimates of the regression coefficients: · · · · · by (19) with $\tau_i = 0$ for Huber ρ function with $c = 1.345$; - - - - - by (19) for Tukey's biweight function with $c = 2$.

If there is a set of positive probability on which $\lambda_{1,k} \to \infty$ along some subsequence, then $\mathcal{L}(B_k, V_k) \to \infty$ since ρ is bounded and $\lambda_{m,k}$ is bounded away from zero. Therefore there must exist an upper bound for $\lambda_{1,k}$.

Suppose $\|B_k\| \to \infty$. We then have

$$\begin{aligned}&\Pr\{(\boldsymbol{y} - B_k\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B_k\boldsymbol{x}) < \tilde{s}\}\\ &\quad \le \Pr\{B_k\boldsymbol{x} = \boldsymbol{0}\} + \Pr\Big\{(\boldsymbol{y} - B_k\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B_k\boldsymbol{x}) < \tilde{s}, \quad B_k\boldsymbol{x} \neq \boldsymbol{0}\Big\},\end{aligned}$$

with the last term going to zero since we know $\lambda_{m,k}$ is bounded away from zero and $\lambda_{1,k}$ is bounded above. This leads to a contradiction to Remark 2.2 about Assumption 2.2(b). The proof of Theorem 2.1 is complete.

Proof of Theorem 2.2. If Assumption 2.2(c) does not hold, then there exists a $B \in \mathcal{R}^{m\times p}$ such that for all $V \in \mathcal{P}_m$

$$\int \rho[(\boldsymbol{y} - B\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y} - B\boldsymbol{x})]dF(\boldsymbol{y}, \boldsymbol{x})$$

$$= \int_{\boldsymbol{y}\neq B\boldsymbol{x}} \rho[(\boldsymbol{y}-B\boldsymbol{x})^T V_k^{-1}(\boldsymbol{y}-B\boldsymbol{x})]dF(\boldsymbol{y},\boldsymbol{x})$$
$$\leq \rho(\infty)\Pr\{\boldsymbol{y}\neq B\boldsymbol{x}\} \leq \rho(\infty)\varepsilon.$$

That is, there exists a B such that constraint (5) holds for all $V \in \mathcal{P}_m$. But, as V approaches a singular matrix, $\log\|V\| \to -\infty$ and so $\mathcal{L}(B,V) \to -\infty$. This completes the proof of the theorem.

Proof of Theorem 3.1. By Theorem 2.1, under the conditions of Theorem 3.1, the CM-estimates exists. Denote by (B,V) the CM-estimates of the original data set.

(a) *Upper bound.* Suppose $n_1 > n\varepsilon$. Consider a sequence of contaminated data set obtained by replacing $\{(\boldsymbol{y}_i,\boldsymbol{x}_i); 1\leq i\leq n_1\}$ by $\{(\boldsymbol{y}_{i,k},\boldsymbol{x}_i); 1\leq i\leq n_1\}$ with $\|\boldsymbol{y}_{i,k}\| \to \infty$ as $k\to\infty$ for $1\leq i\leq n_1$, and let (B_k,V_k) represent the corresponding CM-estimates of the contaminated data sets. Denote the largest and smallest eigenvalues of V_k by $\lambda_{1,k}$ and $\lambda_{m,k}$, respectively. If breakdown does not occur under ε_{n_1}-contamination, then none of (i)−(iii) in the definition of breakdown would happen, that is, there must exist constants a, v_1, v_2 such that $\|B_k\| < a < \infty$ and $0 < v_1 < \lambda_{m,k} \leq \lambda_{1,k} < v_2 < \infty$. Hence, as $k\to\infty$,

$$(\boldsymbol{y}_{i,k}-B_k\boldsymbol{x}_i)^T V_k^{-1}(\boldsymbol{y}_{i,k}-B_k\boldsymbol{x}_i) \to \infty$$

for $i = 1,2,\ldots,n_1$ and so for large k,

$$\liminf_{k\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left[\rho\Big\{(\boldsymbol{y}_{i,k}-B_k\boldsymbol{x}_i)^T V_k^{-1}(\boldsymbol{y}_{i,k}-B_k\boldsymbol{x}_i)\Big\}\right] \geq \frac{n_1\rho(\infty)}{n} > \varepsilon\rho(\infty),$$

This shows that when k is large, the constraint (3) is violated and so $\varepsilon_n^* \leq (\lfloor n\varepsilon\rfloor+1)/n$.

Now suppose $n_1 > n(1-\varepsilon)-C_{3,n}^{(m,p)}$. By definition, for some $B\in\mathcal{R}^{m\times p}$ there exists $C_{3,n}^{(m,p)}$ data points in Z_n so that $\boldsymbol{y}_i = B\boldsymbol{x}_i$, say $1\leq i\leq C_{3,n}^{(m,p)}$ without loss of generality. If we replace n_1 other data points in Z_n by $(\boldsymbol{y}_1,\boldsymbol{x}_1)$, then $\tilde{Z}_n$ has $C_{3,n}^{(m,p)}+n_1$ data points for which $\boldsymbol{y}_i = B\boldsymbol{x}_i$. Since $(C_{3,n}^{(m.p)}+n_1)/n > (1-\varepsilon)$, Theorem 2.2 implies $(\hat{B}(\tilde{Z}_n),\hat{V}(\tilde{Z}_n)) \notin \mathcal{R}^{m\times p}\times\mathcal{P}_m$, which results in one of (i)−(iii). That is, the estimates break down. Therefore $\varepsilon_n^* \leq (\lceil n(1-\varepsilon) - C_{3,n}^{(m,p)}\rceil)/n$.

(b) *Lower bound.* Suppose $n_1 < \min\{\lceil n\varepsilon\rceil, \lceil n(1-\varepsilon)-C_n\rceil\}$. For any ε_{n_1} -contaminated samples $\tilde{Z}_n = \{(\boldsymbol{y}_i^*,\boldsymbol{x}_i^*); 1\leq i\leq n\}$, the parameters $C_{1,n}^{(m,p)}$ and $C_{2,n}^{(m,p)}$ for the contaminate dataset are less than $n(1-\varepsilon)$. Thus, Assumption 2.2(a), (b) and (c) are met for the empirical distribution. By Theorem 2.1, the

CM-estimates $(\hat{B}(\tilde{Z}_n), \hat{V}(\tilde{Z}_n))$ exist. To show that breakdown of CM-estimation has not occurred, we need to show that none of (i)−(iii) can happen. That is, if $\lambda_1(\tilde{Z}_n)$ and $\lambda_m(\tilde{Z}_n)$ represent the largest and smallest eigenvalues of $\hat{V}(\tilde{Z}_n)$, we must show that for all possible ε_{n_1}-contaminated samples $\tilde{Z}_n$, $\lambda_1(\tilde{Z}_n)$ is uniformly bounded above, $\lambda_m(\tilde{Z}_n)$ is uniformly bounded below and $\|\hat{B}(\tilde{Z}_n)\|$ is uniformly bounded above. This will imply $\varepsilon_n^* \geq (\min\{\lceil n\varepsilon\rceil, \lceil n(1-\varepsilon) - C_n\rceil\})/n$.

First we show that $\lambda_m(\tilde{Z}_n)$ is uniformly bounded below. If this is not true, then there are ε_{n_1}-contaminated data sets $\{\tilde{Z}_{nk}\}$ such that the CM-estimates are $\hat{B}_k(\tilde{Z}_{nk}), \hat{V}_k(\tilde{Z}_{nk})$ and $\lambda_m(\tilde{Z}_{nk}) \to 0$ as $k \to \infty$. Since $n_1 < n(1-\varepsilon) - C_n$, there exists $\delta > 0$ such that $n_1 + C_n < n(1-\varepsilon-\delta)$. Choose $\tilde{s}_0$ as in the proof of Theorem 2.1 (recall that $\tilde{s}_0$ depends on ρ and ε only). The constraint (3) for each k implies that the number of $i \leq n$ such that

$$[\tilde{\boldsymbol{y}}_i^* - \hat{B}_k(\tilde{Z}_{nk})\tilde{\boldsymbol{x}}_i^*]^T \hat{V}_k(\tilde{Z}_{nk})^{-1} [\tilde{\boldsymbol{y}}_i^* - \hat{B}_k(\tilde{Z}_{nk})\tilde{\boldsymbol{x}}_i^*] \leq \tilde{s}_0$$

is not less than $n(1-\varepsilon-\delta)$, where $(\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{x}}_i)$ are data points in the contaminated set $\tilde{Z}_{nk}$. Noticing the condition on n_1, we know that there are at least $C_n + 1$ such values from the original data set Z_n. That is,

$$\#\{i : [\boldsymbol{y}_i - \hat{B}_k(\tilde{Z}_{nk})\boldsymbol{x}_i]^T \hat{V}_k(\tilde{Z}_{nk})^{-1} [\boldsymbol{y}_i - \hat{B}_k(\tilde{Z}_{nk})\boldsymbol{x}_i] \leq s_0\} \geq C_n + 1. \tag{26}$$

Consequently,

$$\#\{i : \{\boldsymbol{a}_{mk}^T [\boldsymbol{y}_i - \hat{B}_k(\tilde{Z}_{nk})\boldsymbol{x}_i]\}^2 \leq \lambda_m(\tilde{Z}_{nk}) s_0\} \geq C_n + 1, \tag{27}$$

where $\boldsymbol{a}_{mk}$ is a unit eigenvector of $\hat{V}_k(\tilde{Z}_{nk})$ associated with $\lambda_m(\tilde{Z}_{nk})$. If $\{\hat{B}_k(\tilde{Z}_n)\}$ has a bounded subsequence, then we can select a subsequence $\{k'\}$ such that $\{\boldsymbol{a}_{mk'} \to \boldsymbol{a}$ and $\hat{B}_{k'}(\tilde{Z}_{nk'}) \to B$. Recall that $\lambda_m(\tilde{Z}_{nk'}) \to 0$. Then (27) implies that

$$\#\{i : (\boldsymbol{a}^T(\boldsymbol{y}_i - B\boldsymbol{x}_i)) = 0\} \geq C_n + 1 \geq C_{1,n}^{(m,p)} + 1,$$

which contradicts the definition of $C_{1,n}^{(m,p)}$. If $\|\hat{B}_k(\tilde{Z}_{mk})\| \to \infty$, then we may select a subsequence $\{k'\}$ such that $\boldsymbol{a}_{mk'} \to \boldsymbol{a}$ and $\hat{B}_{k'}(\tilde{Z}_{mk'})/\|\hat{B}_{k'}(\tilde{Z}_{mk'})\| \to B$. Then, (27) implies that

$$\#\{i : (\boldsymbol{a}^T B\boldsymbol{x}_i)) = 0\} \geq C_n + 1 \geq C_{2,n}^{(m,p)} + 1,$$

which contradicts to the definition of $C_{2,n}^{(m,p)}$.

Next we show that $\lambda_1(\tilde{Z}_n)$ is uniformly bounded above over all possible $\tilde{Z}_n$. If there is a sequence of n_1 contaminated data sets $\tilde{Z}_{nk}$ such that $\lambda_m(\tilde{Z}_{nk}) \to \infty$, then, by what has been proved, $\lambda_m(\tilde{Z}_{nk}$ is bounded from below. Thus, we have

$L\{\hat{B}_k(\tilde{Z}_{nk}), \hat{V}_k(\tilde{Z}_{nk}); \tilde{Z}_{nk}\} \to \infty$. Note that $n_1 < \lceil n\varepsilon \rceil$. For any n_1-contaminated data set $\tilde{Z}_{nk}$, for $(B_k, V_k) = (0, \lambda I)$ we have

$$\lim_{\lambda\to\infty} \frac{1}{n} \sum_{i=1}^{n} \rho(\frac{\tilde{\boldsymbol{y}}_i^T \tilde{\boldsymbol{y}}_i}{\lambda}) \le \frac{n_1}{n}\rho(\infty) < \varepsilon\rho(\infty).$$

This shows that there is a λ so that the constraint (3) holds for all $\tilde{Z}_{nk}$, hence that $\hat{B}_k(\tilde{Z}_{nk}), \hat{V}_k(\tilde{Z}_{nk})$ are not CM estimates of the data sets $\tilde{Z}_{nk}$ for all large k. This contradiction completes the proof of the boundedness of $\lambda_1(\tilde{Z}_n)$.

Finally, it is noted that (26) also implies that $\|\hat{B}(\tilde{Z}_n)\|$ is bounded since $\lambda_m(\tilde{Z}_n) \le \lambda_1(\tilde{Z}_n)$, which is bounded above. Otherwise, there must exist $B \neq 0$ so that

$$\#\{i : B\boldsymbol{x}_i = \boldsymbol{0}\} \ge C_n + 1 \ge C_{2,n}^{(m,p)} + 1,$$

which contradicts the definition of $C_{2,n}^{(m.p)}$.

Proof of Theorem 5.1. Without loss of generality, we can assume that $B = 0$, and hence $\boldsymbol{y}_i = \boldsymbol{e}_i$ for any $i \ge 1$. By Assumption 5.3, it can be seen that $\{\tau_n\}$ converges. Hence, we may assume that τ_n is a constant in (19).

Note that $\rho(s)$ is a bounded nondecreasing function for $s \ge 0$. Hence by (19), there exists $M > 0$ such that for any n, $\mathrm{tr}(V_n) < M$, which in turn implies that

$$\max_{n(k)<\ell\le n(k+1)} \|V_l - V_{n(k)}\| = o(1), \quad \text{a.s..} \tag{28}$$

Let $\overline{H}_1(B, V) = \mathrm{E}[H_1(B, V, \boldsymbol{x}, \boldsymbol{y})]$. By (19), we have, for $n(k) < \ell \le n(k+1)$,

$$\begin{aligned} B_\ell &= B_{n(k)} + \sum_{i=n(k)+1}^{\ell} a_{i-1} H_1(B_{i-1}, V_{i-1}, \boldsymbol{x}_i, \boldsymbol{e}_i)\tilde{S}_i^{-1} \\ &= B_{n(k)} + \sum_{i=n(k)+1}^{\ell} i^{-1} a_{i-1} \overline{H}_1(B_{n(k)}, V_{n(k)})A + R_{k,\ell} + T_{k,\ell} + W_{k,\ell}, \end{aligned} \tag{29}$$

where

$$\begin{aligned} R_{k,\ell} &= \sum_{i=n(k)+1}^{\ell} i^{-1} a_{i-1} \Big[H_1(B_{i-1}, V_{i-1}, \boldsymbol{x}_i, \boldsymbol{e}_i) - \overline{H}_1(B_{i-1}, V_{i-1}) \Big] A, \\ T_{k,\ell} &= \sum_{i=n(k)+1}^{\ell} i^{-1} a_{i-1} \Big[\overline{H}_1(B_{i-1}, V_{i-1}) - \overline{H}_1(B_{n(k)}, V_{n(k)}) \Big] A, \\ W_{k,\ell} &= \sum_{i=n(k)+1}^{\ell} i^{-1} a_{i-1} H_1(B_{i-1}, V_{i-1}, \boldsymbol{x}_i, \boldsymbol{e}_i) \Big[(i^{-1}\tilde{S}_i)^{-1} - A \Big]. \end{aligned}$$

Note that ρ is bounded nondecreasing function for $s \geq 0$. By Lemma 2.1 of Bai and Wu (1993), Assumptions 5.2 and 5.4 and (28), it can be shown that

$$\max_{n(k)<\ell\leq n(k+1)} |R_{k,\ell}| = o(d_k); \tag{30}$$

$$\max_{n(k)<\ell\leq n(k+1)} |T_{k,\ell}| \leq O(d_k)\Big(\max_{n(k)<\ell\leq n(k+1)} \|B_\ell - B_{n(k)}\| + o(1)\Big); \tag{31}$$

$$\max_{n(k)<\ell\leq n(k+1)} |W_{k,\ell}| = o(d_k). \tag{32}$$

Hereafter the symbols $o(.)$ and $O(.)$ are in the sense of "with probability one", unless otherwise specified.

Note that by Assumptions 5.1 and 5.4, it can be shown that

$$\sum_{i=n(k)+1}^{\ell} i^{-1} a_{i-1} \overline{H}_1(B_{n(k)}, V_{n(k)}) = O(d_k).$$

Hence, by (29)−(32),

$$\max_{n(k)<\ell\leq n(k+1)} \|B_\ell - B_{n(k)}\| = O(d_k). \tag{33}$$

Since $d_k \to 0$ as $k \to \infty$, to show that $B_n \to 0$, a.s., one need only prove that

$$B_{n(k)} \to 0, \quad \text{a.s.}. \tag{34}$$

By (31) and (33), we have $\max_{n(k)<\ell\leq n(k+1)} |T_{k,\ell}| = o(d_k)$. Therefore,

$$B_{n(k+1)} = B_{n(k)} + \sum_{i=n(k)+1}^{n(k+1)} i^{-1} a_{i-1} \overline{H}_1(B_{n(k)} V_{n(k)}) + o(d_k). \tag{35}$$

Here the estimate $o(d_k)$ holds uniformly for $n(k) < \ell \leq n(k+1)$. Since $B = 0$,

$$\overline{H}_1(B, V) = [\det(\Sigma)]^{-\frac{1}{2}} \int\int h(\|\boldsymbol{e} - B\boldsymbol{x}\|^2_{\tilde{V}})(\boldsymbol{e} - B\boldsymbol{x})\boldsymbol{x}^T f_1(\|\boldsymbol{e}\|_\Sigma) d\boldsymbol{e} dF_2(\boldsymbol{x}),$$

and then

$$\begin{aligned}\mathrm{tr}[B^T \overline{H}_1(B, V)] = \mathrm{tr}[\det(\Sigma)]^{-\frac{1}{2}} \int\int (B\boldsymbol{x})^T h(\|\boldsymbol{e} - B\boldsymbol{x}\|^2_{\tilde{V}})(\boldsymbol{y} - B\boldsymbol{x}) \\ \times f_1(\|\boldsymbol{e}\|_\Sigma) d\boldsymbol{e} dF_2(\boldsymbol{x})].\end{aligned}$$

Let $\boldsymbol{z} = O_2^T O_1^T \Sigma^{-1/2} \boldsymbol{e}$, $\tilde{B} = O_2^T O_1^T \Sigma^{-1/2} B$, where O_1 and O_2 are orthogonal matrices such that $\Delta_1 = O_1^T \Sigma^{-1/2} \tilde{V} \Sigma^{-1/2} O_1$ and $\Delta_2 = O_2^T O_1^T \Sigma O_1 O_2$ are diagonal. Then we have

$$\mathrm{tr}[\det(\Sigma)]^{-\frac{1}{2}} \int (B\boldsymbol{x})^T h(\|\boldsymbol{e} - B\boldsymbol{x}\|^2_{\tilde{V}})(\boldsymbol{e} - B\boldsymbol{x}) f_1(\|\boldsymbol{z}\|) d\boldsymbol{e}$$

$$= \mathrm{tr} \int (\tilde{B}\boldsymbol{x})^T \Delta_2 h(\|\boldsymbol{z} - \tilde{B}\boldsymbol{x}\|^2_{\Delta_1})(\boldsymbol{z} - B\boldsymbol{x}) f_1(\|\boldsymbol{z}\|) d\boldsymbol{e}.$$

Mimicking the proof of (3.12) in Bai and Wu (1993), it can be shown that Condition (ii) of Lemma 2.2 there is satisfied. Hence by applying Lemma 2.2 of Bai and Wu (1993), we have that $B_{n(k)} \to 0$, a.s., which in turn implies that $B_n \to 0$, a.s..

Further assume that Assumption 5.5 holds. By similar arguments as in the proof of Theorem 3.2 in Bai and Wu (1993), it can be shown that $V_n \to \Omega$, a.s..

Acknowledgement

We thank the referees for helpful comments and suggestions. We also thank Prof. Jiahua Chen for reviewing an earlier draft of this paper. The research was supported by the Natural Sciences and Engineering Research Council of Canada. The work of Prof. Bai was supported by NSFC Grant 10571020 and NUS Grant R-155-000-061-112.

References

Bai, Z. and Wu, Y. (1993). Recursive algorithm for M-estimators of regression coefficients and scatter parameters in linear models. *Sankhyā Ser. B* **55**, 199-218.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428-434.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (Edited by P. J. Bickel, K. A. Doksum and J. L. Hodges), 157-184. Wadsworth, Belmont, CA.

Englund, J.-E., Holst, U. and Ruppert, D. (1988). Recursive M-estimate of location and scale for dependent sequences. *Scand. J. Statist.* **15**, 147-159.

Englund, J.-E. (1993). Multivariate recursive M-estimate of location and scatter for dependent sequences. *J. Multivariate Anal.* **45**, 257-273.

Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Ann. Statist.* **24**, 1346-1370.

Lopuhaä, H. P. (1989). On the relationship between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.*, **17**, 1662-1683.

Mendes, B. and Tyler, D. E. (1996). Constrained M-estimation for regression. In *Robust Statistics, Data Analysis, and Computer Intensive Methods* **109** (Edited by H. Rieder), 299-320. Springer-Verlag, New York.

Miao, B. Q. and Wu, Y. (1996). Limiting behavior of recursive M-estimate in multivariate linear regression models. *J. Multivariate Anal.* **59**, 60-80.

Rubinstein, R. Y. (1986). *Monte Carlo Optimization, Simulation, and Sensitivity of Queuing Networks.* Wiley, New York.

Wu, Y. (1996). On consistency of recursive multivariate M-estimators in linear models. In *Robust Statistics, Data Analysis, and Computer Intensive Methods* **109** (Edited by H. Rieder), 411-424. Springer-Verlag, New York.

School of Mathematics & Statistics, Northeast Normal University, Changchun, Jilin, China.

E-mail: baizd@nenu.edu.cn

Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 1, Singapore 117546.

E-mail: stabaizd@nus.edu.sg

Department of Mathematics, Graduate University of the Chinese Academy of Sciences, Beijing, China.

Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada.

E-mail: wuyh@mathstat.yorku.ca

(Received July 2006; accepted January 2007)