Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 37475, 15 pages
doi:10.1155/2007/37475

Research Article

An M-Estimation-Based Procedure for Determining the Number of Regression Models in Regression Clustering

C. R. Rao, Y. Wu, and Q. Shao

Received 16 June 2007; Accepted 16 July 2007

Recommended by Paul Cowpertwait

In this paper, a procedure based on M-estimation to determine the number of regression models for the problem of regression clustering is proposed. We have shown that the true classification is attained when n increases to infinity under certain mild conditions, for instance, without assuming normality of the distribution of the random errors in each regression model.

1. Introduction

Cluster analysis is a statistical tool to classify a set of objects into groups so that objects within a group are "similar" and objects in different groups are "dissimilar." The purpose of clustering is to discover "natural" structure hidden in a data set. Regression clustering technique is often used to classify the data and recover the underlying structure, when the data set is believed to be a random sample from a population comprising of a fixed, but unknown, number of subpopulations, each of which is characterized by a distinct regression model. Regression clustering is one of the most commonly used model-based clustering techniques. It has been studied by Bock [1], Quandt and Ramsey [2], and Späth [3] among others, and has applications in a variety of disciplines, for example, in market segmentation by DeSarbo and Cron [4] and quality control systems by Lou et al. [5].

A fundamental problem, as well as a preliminary step in regression clustering, is to determine the underlying "true" number of regression models in a data set. Shao and Wu [6] proposed an information-based criterion (named criterion "LS-C" in the sequel) to tackle this problem. The limiting behavior of LS-C is given in their paper.

However, it is well known that the least squares (LS) method is very sensitive to outliers and violation of the normality assumption of the data. This instability also exists in the

LS-based procedures for both selecting the number of regression models and classifying the data in the context of regression clustering.

During the past three decades, numerous efforts have been made for developing robust statistical procedures for statistical inferences. Among them, procedures based on M-estimators, which are maximum likelihood-type estimators (Hampel et al. [7] and Huber [8]), play an important role. The M-estimation-based model selection criteria are considered by Konishi and Kitagawa [9], Machado [10], and Wu and Zen [11] among others.

To overcome the instability of the LS-based procedures in regression clustering, we propose an M-estimation-based procedure for determining the number of regression models, which is an extension of M-estimation-based information criterion for linear model selection developed by Wu and Zen [11]. Its asymptotic behavior will be investigated.

The structure of this paper is arranged as follows. In Section 2, we build a probabilistic framework for our problem and introduce some notations used in this paper. Section 3 lists all the assumptions needed for our study. In Section 4, we study the limiting behavior of the proposed criterion. Some ancillary results required for our proofs are presented in the appendix.

2. Notation and preliminaries

We consider the clustering problem for n objects $\mathbb{O}^{(n)} = \{1,\ldots,n\}$, where for each object j, $(\mathbf{x}_j, y_j)$ has been recorded, where $\mathbf{x}_j = (x_j^{(1)},\ldots,x_j^{(p)})' \in \mathbb{R}^p$ is a nonrandom explanatory p-vector and $y_j \in \mathbb{R}$ is a random response variable. The set of these n objects is a random sample from a structured population as specified below.

Suppose that there exists an underlying partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1^{(n)},\ldots,\mathbb{O}_{k_0}^{(n)}\}$ for these n objects, and each component $\mathbb{O}_i^{(n)} \triangleq \{i_1,\ldots,i_{n_i}\} \subseteq \mathbb{O}^{(n)}$ is characterized by a linear regression model:

$$y_{j,\mathbb{O}_i} = \mathbf{x}'_{j,\mathbb{O}_i}\boldsymbol{\beta}_{0i} + e_{j,\mathbb{O}_i}, \quad j = i_1,\ldots,i_{n_i}, \tag{2.1}$$

where $n_i = |\mathbb{O}_i|$ is the number of observations in the ith component $\mathbb{O}_i$, $i = 1,\ldots,k_0$, and $\sum_{i=1}^{k_0} n_i = n$. Note that $\mathbb{O}_i$ and $\mathbb{O}_i^{(n)}$ are used interchangeably to denote the ith component of the underlying partition $\Pi_{k_0}^{(n)}$. $(\mathbf{x}_{j,\mathbb{O}_i}, y_{j,\mathbb{O}_i})$ $(j = i_1,\ldots,i_{n_i},\ i = 1,\ldots,k_0)$ is a relabeled observation $(\mathbf{x}_j, y_j)$ $(j = 1,\ldots,n)$ to represent the jth object in the ith component $\mathbb{O}_i$ of the true partition $\Pi_{k_0}^{(n)}$. We will use this double-index notation for any object $(\mathbf{x}_j, y_j)$ throughout this paper to identify the component to which it belongs. $\boldsymbol{\beta}_{0i} \in \mathbb{R}^p$ are k_0 pairwise distinct p-vectors of unknown regression parameters, and $e_{j,\mathbb{O}_i}$, $j = i_1,\ldots,i_{n_i}$, are independently and identically distributed random errors for $i = 1,\ldots,k_0$.

However, this underlying structure (2.1) is not observable. What we observe is just a random sample of n objects with the data values $(\mathbf{x}_j, y_j)$ for each of the $p+1$ variables associated with each object. Our task is then to reconstruct the hidden structure (2.1) from the observed data by first estimating the number of regression models k_0 and then classifying the data and estimating the regression parameters in each regression model accordingly.

Consider any possible classification of these n objects: $\Pi_k^{(n)} = \{\mathscr{C}_1^{(n)},\ldots,\mathscr{C}_k^{(n)}\}$, where $k \le K$ is a positive integer. For this partitioning, we fit k M-estimator-based linear regression models and obtain kM-estimates $\widehat{\boldsymbol{\beta}}_s$, $s = 1,\ldots,k$, separately. Then the M-estimator-based criterion for estimating the number of regression models is given as follows: let $q(k)$ be a strictly increasing function of k and let A_n be a sequence of constants. We define

$$R_n\left(\Pi_k^{(n)}\right) = \sum_{s=1}^{k} \sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\widehat{\boldsymbol{\beta}}_s\right) + q(k)A_n, \tag{2.2}$$

where ρ is a convex discrepancy function. As an example, ρ can be chosen as Huber's discrepancy function

$$\rho_c(t) = \begin{cases} \dfrac{1}{2}t^2, & |t| < c, \\ c|t| - \dfrac{1}{2}c^2, & |t| \ge c. \end{cases} \tag{2.3}$$

Also, in (2.2), $\sum_{j\in\mathscr{C}_s}$ stands for the summation made over all the observations in the class $\mathscr{C}_s$ and $\widehat{\boldsymbol{\beta}}_s$ is the M-estimator in the sth class such that

$$\sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\widehat{\boldsymbol{\beta}}_s\right) = \min_{\boldsymbol{\beta}_s} \sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\boldsymbol{\beta}_s\right). \tag{2.4}$$

Again, $\mathscr{C}_s$ and $\mathscr{C}_s^{(n)}$ are used interchangeably in the above equations to denote the sth class in the partition $\Pi_k^{(n)}$. We will continue this convenient usage without further explanation in the sequel. It can be seen that in (2.2), the first term is a generalization of a minimum negative log-likelihood function and the second term is the penalty for over-fitting.

Then the estimate of the underlying number of regression models k_0, $\widehat{k}_n$ is obtained by minimizing the criterion (2.2), that is,

$$\widehat{k}_n = \arg\min_{1\le k\le K} \min_{\Pi_k^{(n)}} R_n\left(\Pi_k^{(n)}\right). \tag{2.5}$$

We will call this criterion MR-C, which stands for the M-estimator-based regression clustering. Moreover, criterion MR-C in (2.5) shows that it actually determines the optimal number of regression models and the associated partitioning simultaneously.

3. Assumptions

Let $\mathbb{O}_l = \{l_1,\ldots,l_{n_l}\}$ be any component or a subset of a component associated with the underlying true partition $\Pi_{k_0}^{(n)}$ of $\mathbb{O}^{(n)}$, and $n_l = |\mathbb{O}_l|$. If we let $X_{n_l} = (\mathbf{x}_{l_1,\mathbb{O}_l},\ldots,\mathbf{x}_{l_{n_l},\mathbb{O}_l})'$ be the design matrix in $\mathbb{O}_l$, then $W_{n_l} = X'_{n_l}X_{n_l}$, $d_{n_l}^2 = \max_{1\le j\le n_l} \mathbf{x}'_{j,\mathbb{O}_l} W_{n_l}^{-1}\mathbf{x}_{j,\mathbb{O}_l}$.

To facilitate the study on the limiting behavior of the criterion MR-C, we need the following assumptions.

(A) For the true partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1,\ldots,\mathbb{O}_{k_0}\}$ and $n_i = |\mathbb{O}_i|$, there exists a fixed constant $a_0 > 0$ such that

$$a_0 n \le n_i \le n \quad \forall\, i = 1,\ldots,k_0. \tag{3.1}$$

Remark 3.1. This assumption is equivalent to the explicit assumption that the population comprises k_0 subpopulations with proportions $\pi_1,\ldots,\pi_{k_0}$ where $0<\pi_i<1$, $i=1,\ldots,k_0$, $\sum_{i=1}^{k_0}\pi_i=1$. Then $a_0=\min_{1\le i\le k_0}\pi_i$ would satisfy (3.1).

(B1) $\rho(\cdot)$ is a convex function on $\mathbb{R}^1$.

(B2) $\mathrm{E}[\rho(e_{j,\mathbb{O}_i})]$ is finite for all $j\in\mathbb{O}_i$ and $i=1,\ldots,k_0$.

(B3) For any $\boldsymbol{\beta}$ and observations in $\mathbb{O}_l$,

$$\liminf_{n_l\to\infty}\frac{1}{n_l}\sum_{j\in\mathbb{O}_l}\mathrm{E}\left[\rho\left(e_{j,\mathbb{O}_l}-\mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}\right)-\rho\left(e_{j,\mathbb{O}_l}\right)\right]\ge g(\boldsymbol{\beta}),\tag{3.2}$$

where $g(\cdot)$ is a nonnegative convex function and is strictly convex in a neighborhood of $\mathbf{0}$.

If ρ has a first-order derivative, in order to find M-estimator of $\boldsymbol{\beta}_s$ in the sth-class, one may first find all first-order partial derivatives of $\sum_{j\in\mathbb{O}_s}\rho(y_{j,\mathbb{O}_s}-\mathbf{x}'_{j,\mathbb{O}_s}\boldsymbol{\beta}_s)$ and then set them to be equal to zeros. The simultaneous solutions of these equations give the M-estimator of $\boldsymbol{\beta}_s$. However in some cases, ρ does not have a first-order derivative. Note that for any convex function, it always has subgradients, which are just partial derivatives if they do exist (see Rockafellar [12]). Let $\psi(\cdot)$ be any choice of the subgradient of $\rho(\cdot)$ and denote by $\mathcal{U}$ the set of discontinuity points of ψ, which is the same for all choices of ψ.

(C1) The common distribution function F of $e_{j,\mathbb{O}_i}$, $j\in\mathbb{O}_i$, is unimodal and satisfies $F(\mathcal{U})=0$. $\mathrm{E}[\psi(e_{j,\mathbb{O}_i})]=0$, $\mathrm{E}[\psi^2(e_{j,\mathbb{O}_i})]=\sigma_i^2<\infty$ for any $i=1,\ldots,k_0$, and

$$\mathrm{E}\left[\psi\left(e_{j,\mathbb{O}_i}+u\right)\right]=a_iu+o(|u|),\quad \text{as } u\longrightarrow 0,\tag{3.3}$$

where a_i, $i=1,\ldots,k_0$, are finite positive constants.

(C2) There exist positive constants ζ and h_0 such that for any $h\in[0,h_0]$ and any u,

$$\psi(u+h)-\psi(u)\le\zeta.\tag{3.4}$$

(C3) The moment generating function $M_i(t)=\mathrm{E}[\exp\{t\psi(e_{j,\mathbb{O}_i})\}]$ exists for $|t|\le\Delta$, where $i=1,\ldots,k_0$.

(C4) $\mathrm{E}[|\psi(e_{j,\mathbb{O}_i})|^3]<\infty$, $j\in\mathbb{O}_i$, $i=1,\ldots,k_0$.

Denote the eigenvalues of a symmetric matrix B of order p by $\lambda_1(B)\ge\cdots\ge\lambda_p(B)$.

(X) There are constants a_1 and a_2 such that

$$0<a_1n_l\le\lambda_p\left(W_{n_l}\right)\le\lambda_1\left(W_{n_l}\right)\le a_2n_l\quad\text{for large enough } n_l.\tag{3.5}$$

The following three assumptions are on d_{n_l}. Recall that $d_{n_l}^2=\max_{1\le j\le n_l}\mathbf{x}'_{j,\mathbb{O}_l}W_{n_l}^{-1}\mathbf{x}_{j,\mathbb{O}_l}$.

(X1) $d_{n_l}(\log\log n_l)^{1/2}\to 0$ as $n_l\to\infty$.

(X2) $d_{n_l}(\log n_l)^{1+\iota}=O(1)$, where $\iota>0$ is a constant.

(X3) When n_l is large enough, there exists a constant $\omega>0$ such that $d_{n_l}\le\omega n_l^{-1/2}$.

Remark 3.2. Assumptions (X) and (X1)–(X3) describe essentially the behavior of the explanatory variables. Assumptions (X1)–(X3) are imposed so that d_{n_l} converges to 0 at certain rates. It can be seen that Assumption (X) is satisfied almost surely if $\mathbf{x}_i$, $i=1,2,\ldots$, are independently and identically distributed observations of a random vector $\mathbf{X}$ with

strictly positive definite covariance matrix. If we further assume that $|\mathbf{X}|$ is finite, then (X1)–(X3) are met almost surely.

(Z) The sequence $\{A_n\}$ satisfies

$$\frac{A_n}{n} \longrightarrow 0, \qquad \frac{A_n}{\log\log n} \longrightarrow \infty. \tag{3.6}$$

Excluding Assumption (A), all other assumptions are ordinarily used in the study of limiting behavior of an M-estimator. The only difference is that we now require them to hold in any sth-class, $1 \le s \le k$.

4. Limiting behavior of the criterion MR-C

Suppose that (B1)–(B3), (C1)–(C3), (X), (X1), and (Z) hold.

Let $\Pi_{k_0}^{(n)}$ be the true underlying partition of the n objects with the model structure (2.1). Observe that the true partition $\Pi_{k_0}^{(n)}$ is a sequence of naturally nested classifications as n increases, that is,

$$\mathbb{O}_i^{(n)} \subseteq \mathbb{O}_i^{(n+1)}, \quad i = 1,\ldots,k_0, \text{ for large } n. \tag{4.1}$$

Consider a given sequence of classifications with k clusters $\Pi_k^{(n)} = \{\mathcal{C}_1^{(n)},\ldots,\mathcal{C}_k^{(n)}\}$ of $\mathbb{O}^{(n)}$ such that

$$\mathcal{C}_s^{(n)} \subseteq \mathcal{C}_s^{(n+1)}, \quad s = 1,\ldots,k, \text{ for large } n, \tag{4.2}$$

when n increases. For simplicity, when no confusion appears, n will be suppressed in $\Pi_{k_0}^{(n)}$, $\Pi_k^{(n)}$, $\mathbb{O}_i^{(n)}$, $1 \le i \le k_0$, and $\mathcal{C}_s^{(n)}$, $1 \le s \le k$.

Consider the following two cases.

Case 1. $k_0 < k < K$, where $K < \infty$ is a fixed constant:

$$\begin{aligned} &R_n(\Pi_k^{(n)}) - R_n(\Pi_{k_0}^{(n)}) \\ &\quad = \sum_{s=1}^{k} \sum_{j\in\mathcal{C}_s} \rho(y_{j,\mathcal{C}_s} - \mathbf{x}'_{j,\mathcal{C}_s}\widehat{\boldsymbol{\beta}}_s) - \sum_{i=1}^{k_0} \sum_{j\in\mathbb{O}_i} \rho(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}) + (q(k) - q(k_0))A_n, \end{aligned} \tag{4.3}$$

where

$$\widehat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathcal{C}_s} \rho(y_{j,\mathcal{C}_s} - \mathbf{x}'_{j,\mathcal{C}_s}\boldsymbol{\beta}), \tag{4.4}$$

$$\widehat{\boldsymbol{\beta}}_{0i} = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathbb{O}_i} \rho(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\boldsymbol{\beta}). \tag{4.5}$$

Since we have $k_0 < k < K < \infty$, the number of possible intersection sets $\mathcal{C}_s \cap \mathbb{O}_i$ is finite, and

$$\mathbb{O}^{(n)} = \cup_{i=1}^{k_0}\mathbb{O}_i = \cup_{s=1}^{k}\mathcal{C}_s = \cup_{s=1}^{k}\cup_{i=1}^{k_0}(\mathcal{C}_s \cap \mathbb{O}_i). \tag{4.6}$$

Hence

$$\begin{aligned}
&R_n(\Pi_k^{(n)}) - R_n(\Pi_{k_0}^{(n)}) \\
&= \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] + (q(k) - q(k_0))A_n \\
&= \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \\
&\quad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\quad + (q(k) - q(k_0))A_n,
\end{aligned} \tag{4.7}$$

where $\widehat{\boldsymbol{\beta}}_{0si}$ is the M-estimator defined by

$$\widehat{\boldsymbol{\beta}}_{0si} = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\boldsymbol{\beta}\right). \tag{4.8}$$

By (4.4) and (4.8), we have

$$\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \geq 0. \tag{4.9}$$

By (A.3) of Lemma A.2, (4.5), (4.8), and the fact that $\mathscr{C}_s \cap \mathbb{O}_i$ is a subset of the true class $\mathbb{O}_i$, we have that

$$\begin{aligned}
\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right)\right] &= O(\log\log n), \\
\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\left[\rho\left(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right) - \rho\left(e_{j,\mathbb{O}_i}\right)\right] &= O(\log\log n).
\end{aligned} \tag{4.10}$$

Note that $\bigcup_{s=1}^{k}\bigcup_{i\in k_0}\bigcup_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\{e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\}$ is the same as $\bigcup_{i=1}^{k_0}\bigcup_{j\in\mathbb{O}_i}\{e_{j,\mathbb{O}_i}\}$. Hence we have that

$$\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\rho\left(e_{j,\mathbb{O}_i}\right) \equiv \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right). \tag{4.11}$$

We further have

$$\begin{aligned}
&\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)-\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right]\\
&\quad=\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)-\rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right)\right]\\
&\quad\quad-\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\left[\rho\left(y_{j,\mathbb{O}_i}-\mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)-\rho\left(e_{j,\mathbb{O}_i}\right)\right]=O(\log\log n).
\end{aligned}\tag{4.12}$$

Therefore, by (4.9), (4.12), Assumption (Z), and the fact that $q(k)-q(k_0)>0$, we obtain that

$$R_n\left(\Pi_k^{(n)}\right)-R_n\left(\Pi_{k_0}^{(n)}\right)>0,\quad \text{a.s.}\tag{4.13}$$

for n large enough.

Case 2. $k<k_0$.

By [6, Lemma 3.1] for any partition $\Pi_k^{(n)}=\{\mathscr{C}_1,\ldots,\mathscr{C}_k\}$ of $\mathbb{O}^{(n)}$, there exist one class in $\Pi_k^{(n)}$ and two distinct components in the true partition $\Pi_{k_0}^{(n)}=\{\mathbb{O}_1,\ldots,\mathbb{O}_{k_0}\}$, say $\mathscr{C}_1\in\Pi_k^{(n)}$ and $\mathbb{O}_1,\mathbb{O}_2\in\Pi_{k_0}^{(n)}$ such that

$$b_0 n<\left|\mathscr{C}_1\cap\mathbb{O}_1\right|<n,\qquad b_0 n<\left|\mathscr{C}_1\cap\mathbb{O}_2\right|<n,\tag{4.14}$$

where $b_0=a_0/k_0>0$ is a constant.

Let $d_0=\min_{1\le i\ne l\le k_0}|\boldsymbol{\beta}_{0i}-\boldsymbol{\beta}_{0l}|$. Then $d_0>0$ is a fixed constant. Consider

$$\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right),\qquad \sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widehat{\boldsymbol{\beta}}_1\right),\tag{4.15}$$

where $\widehat{\boldsymbol{\beta}}_1$ is the M-estimator in $\mathscr{C}_1$ defined in (4.4) with $s=1$. Then in view of the convexity of $\rho(\cdot)$, by (4.4), (4.14), and the fact that $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{02}$ are two distinct underlying true parameter vectors in the model structure (2.1), at least one of the following two inequalities must hold:

$$\begin{aligned}
&\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right)\\
&\quad>\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widetilde{\boldsymbol{\beta}}\right)\quad\forall\widetilde{\boldsymbol{\beta}}:\left|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_{01}\right|\le\frac{d_0}{4},
\end{aligned}\tag{4.16}$$

$$\begin{aligned}
&\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widehat{\boldsymbol{\beta}}_1\right)\\
&\quad>\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widetilde{\boldsymbol{\beta}}\right),\quad\forall\widetilde{\boldsymbol{\beta}}:\left|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_{02}\right|\le\frac{d_0}{4}.
\end{aligned}\tag{4.17}$$

Without loss of generality, we assume that (4.16) holds. Now let us focus our discussion on the set $\mathscr{C}_1 \cap \mathbb{C}_1$. Let $n_{11} = |\mathscr{C}_1 \cap \mathbb{C}_1|$ be the number of objects in the set $\mathscr{C}_1 \cap \mathbb{C}_1$. We intend to find the order of

$$\sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_1) - \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011})] \stackrel{\text{def}}{=} T \tag{4.18}$$

in terms of n as n increases to infinity. In the above expression for T, $\widehat{\boldsymbol{\beta}}_{01}$ is the M-estimator in $\mathbb{C}_1$ defined in (4.5) with $i = 1$ and $\widehat{\boldsymbol{\beta}}_{011}$ is the M-estimator in $\mathscr{C}_1 \cap \mathbb{C}_1$ defined as follows:

$$\widehat{\boldsymbol{\beta}}_{011} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \boldsymbol{\beta}). \tag{4.19}$$

$\mathscr{C}_1 \cap \mathbb{C}_1$ is a subset of $\mathbb{C}_1 \in \Pi_{k_0}^{(n)}$ which is the underlying true classification of $\mathbb{C}^{(n)}$. By (A.4), Lemma A.2, with probability one, $|\widehat{\boldsymbol{\beta}}_{011} - \boldsymbol{\beta}_{01}| < d_0/4$ for n_{11} large enough, where $\widehat{\boldsymbol{\beta}}_{011}$ is defined in (4.19). Let $\overline{D} \stackrel{\text{def}}{=} \{\widetilde{\boldsymbol{\beta}} : |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{01}| = d_0/4\}$. Then by (4.16), it is certain that there exists a point $\widetilde{\boldsymbol{\beta}}_{\overline{D}} \in \overline{D}$ such that

$$\sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_1) > \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}). \tag{4.20}$$

Hence

$$\begin{aligned} T &> \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}) - \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011})] \\ &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))] \\ &\quad - \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))] \stackrel{\text{def}}{=} T_1 + T_2. \end{aligned} \tag{4.21}$$

By (A.6), Lemma A.3, there exists a constant $\delta > 0$ such that

$$T_1 \geq \delta n_{11} + o(n_{11}), \quad \text{a.s.} \tag{4.22}$$

Write $T_2 = T_{21} + T_{22}$ with

$$\begin{aligned} T_{21} &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011}) - \rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1})], \\ T_{22} &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))]. \end{aligned} \tag{4.23}$$

By (A.3), Lemma A.2, and (4.2), we have

$$T_{21} = O(\log\log n_{11}), \quad \text{a.s.} \tag{4.24}$$

By (B2), (4.2), and the strong law of large numbers, we obtain

$$T_{22} = o(n_{11}), \quad \text{a.s.} \tag{4.25}$$

Hence, by (4.24) and (4.25), we have

$$T_2 = o(n_{11}), \quad \text{a.s.} \tag{4.26}$$

In view of (4.21), (4.22), and (4.26), it follows that

$$\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1} \left[\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right) - \rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_{011}\right)\right] > \delta n_{11} + o(n_{11}), \quad \text{a.s.} \tag{4.27}$$

By (4.9), we can express our object function as follows:

$$\begin{aligned}
&R_n\left(\Pi_k^{(n)}\right) - R_n\left(\Pi_{k_0}^{(n)}\right) \\
&\quad = \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \\
&\qquad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad + \left(q(k) - q(k_0)\right)A_n \\
&\quad \geq \sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1} \left[\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right) - \rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_{011}\right)\right] \\
&\qquad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad + \left(q(k) - q(k_0)\right)A_n,
\end{aligned} \tag{4.28}$$

where $\widehat{\boldsymbol{\beta}}_s$, $1 \leq s \leq k$, and $\widehat{\boldsymbol{\beta}}_{0i}$, $1 \leq i \leq k_0$, are defined in (4.4) and (4.5). By the same argument as used in Case 1, we have

$$\begin{aligned}
&\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad = O(\log\log n) = o(n).
\end{aligned} \tag{4.29}$$

Therefore, by (3.6), (4.14), (4.27), and (4.29), we have

$$\begin{aligned}
&R_n\left(\Pi_k^{(n)}\right) - R_n\left(\Pi_{k_0}^{(n)}\right) \\
&\qquad > \delta n_{11} + o(n_{11}) + o(n) + \left[q(k) - q(k_0)\right]A_n \\
&\qquad \geq \delta b_0 n + o(n) + \left[q(k) - q(k_0)\right]A_n > 0, \quad \text{a.s.}
\end{aligned} \tag{4.30}$$

for n large enough.

Table 5.1. Parameter values used in the simulation study of regression clustering.

Case	k_0	Regression coefficients	No. of obs.
1	1	$\boldsymbol{\beta}_0 = \binom{1}{6}$	$n = 120$
2	2	$\boldsymbol{\beta}_{01} = \binom{20}{9}, \boldsymbol{\beta}_{02} = \binom{1}{6}$	$n_1 = 70$ $n_2 = 50$
3	3	$\boldsymbol{\beta}_{01} = \binom{30}{9}, \boldsymbol{\beta}_{02} = \binom{12}{8}, \boldsymbol{\beta}_{03} = \binom{-2}{9}$	$n_1 = 35$ $n_2 = 35$ $n_3 = 50$

Therefore, combining the results from (4.13) in Case 1 and (4.30) in Case 2, we have showed that the true classification is attained when n increases to infinity.

Remark 4.1. In the above discussion, the set of the conditions (B1)–(B3), (C1)–(C3), (X), (X1), and (Z) can be replaced by any set of the following conditions:

(a) (A), (B1)–(B3), (C1)-(C2), (C4), (X), (X2), and (Z);

(b) (A), (B1)-(B2), (C1)-(C2), (C4), (X), (X3), and (Z);

(c) (A), (B1)-(B2), (C1)–(C3), (X), (X3), and (Z).

Remark 4.2. Hannan and Quinn [13] show that $A_n = c\log\log n$ is sufficient for strong consistency in a classical estimation procedure for the order of an autoregression. By computing the upper bound in our proofs carefully, we can show that $A_n = c\log\log n$ also works here.

Remark 4.3. The above study is not feasible when all possible classifications are considered simultaneously. For simplicity, we consider the quadratic ρ function, that is, $\rho(t) = t^2$. Let $D_n = \{\text{all nonempty subsets of } \mathbb{O}\}$, then for any $l \in \{1, 2, \ldots, p\}$,

$$\max_{d\in D_n}\left|\sum_{j\in d} x_j^{(l)} e_j\right| \geq \max\left(\sum_{j:x_j^{(l)}e_j>0} x_j^{(l)} e_j, \sum_{j:x_j^{(l)}e_j<0} (-x_j^{(l)} e_j)\right) \geq \frac{1}{2}\sum_{j=1}^{n}\left|x_j^{(l)} e_j\right|. \tag{4.31}$$

Note that in general, $\sum_{j=1}^{n} |x_j^{(l)} e_j| = O(n)$ for any $l \in \{1, 2, \ldots, p\}$. Hence the key equation (A.2), Lemma A.2 does not hold uniformly for all possible subsets of $\mathbb{O} = \{1, 2, \ldots, n\}$.

5. A simulation study

In this section, we present a simulation study for the finite sample performance of the criterion MR-C. In this simulation, $q(k) = 3k(p + 3)$, where p is the known number of regression coefficients in the model structure (2.1) and k is the number of clusters we consider. Since $\lim_{t\to 0}[(\log n)^t - 1]/t = \log\log n$ holds, by Remark 4.2 in Section 4, we let $A_n^{(i)} = (1/\lambda_i)((\log n)^{\lambda_i}) - 1$, with $\lambda_1 = 1.6$, $\lambda_2 = 1.8$, $\lambda_3 = 2.0$, and $\lambda_4 = 2.2$ employed in the simulation.

We consider one cluster, two cluster and three cluster cases, respectively. In all cases, the covariate is generated from $N(0, 1)$. The parameter values used for each case are given in Table 5.1. $N(0, 1)$ and Cauchy(0, 1) random error terms are used to generate the data

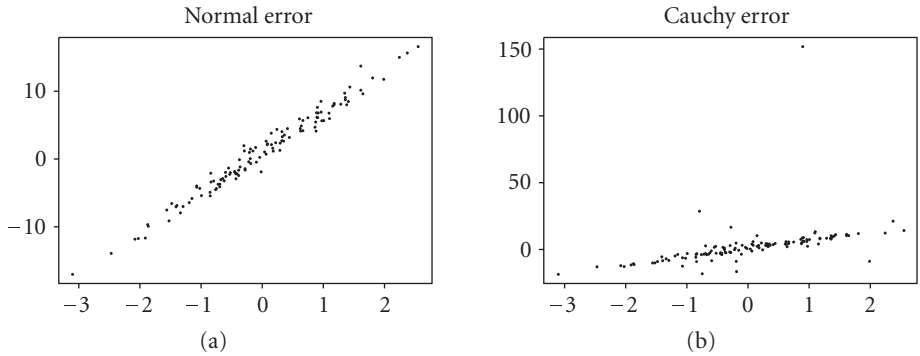

FIGURE 5.1. Plots of simulated data for one homogeneous cluster.

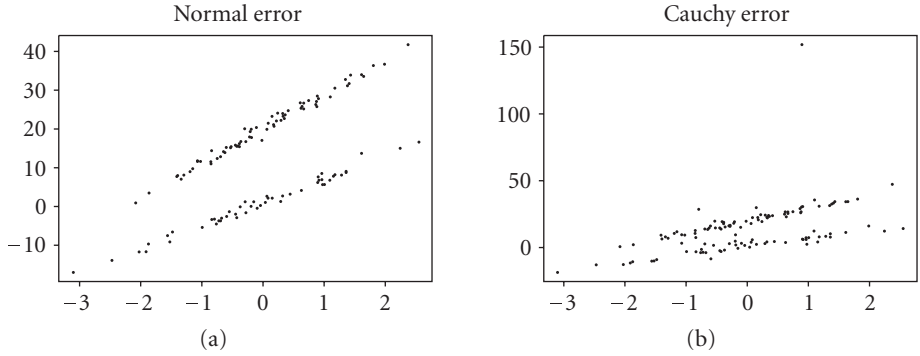

FIGURE 5.2. Plots of simulated data for two separated linear patterns.

for each of the above three cases, respectively. Therefore, in all, we actually consider six models. We use a shorthand notation to identify them:

(i) N1C1 (C1C1) Case 1, one single line, normal (Cauchy) errors;

(ii) N1C2 (C1C2) Case 2, two separated lines, normal (Cauchy) errors;

(iii) N1C3 (C1C3) Case 3, three separated lines, normal (Cauchy) errors.

The ρ functions we employed for M-estimator are (1) $\rho_1(u) = u^2$ (LS); (2) $\rho_2(u) = 0.5u^2$ if $|u| \leq 1.345$ and $\rho_2(u) = 1.345|u| - 0.5 \times 1.345^2$ otherwise (Huber ρ). When ρ is the quadratic discrepancy function, MR-C coincides with LS-C. In the following, MR-C stands for the M-estimator-based regression clustering procedure with Huber's ρ. In order to keep the same scale between LS-C and MR-C, the actual LS-C implemented in this simulation study is to minimize $\sum_{s=1}^{k} \sum_{j \in \mathscr{C}_s} (y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s} \widehat{\boldsymbol{\beta}}_s)^2/2 + q(k)A_n$ over all possible partitions. It is clear that this slight modification does not affect the asymptotic property of LS-C.

Figures 5.1, 5.2, and 5.3 give us an intuitive idea of what the data look like for Cases 1, 2, and 3 with $N(0,1)$ and Cauchy(0, 1) errors, respectively. These figures show that the groupings of linear patterns are quite apparent and clear in each case for the normal error

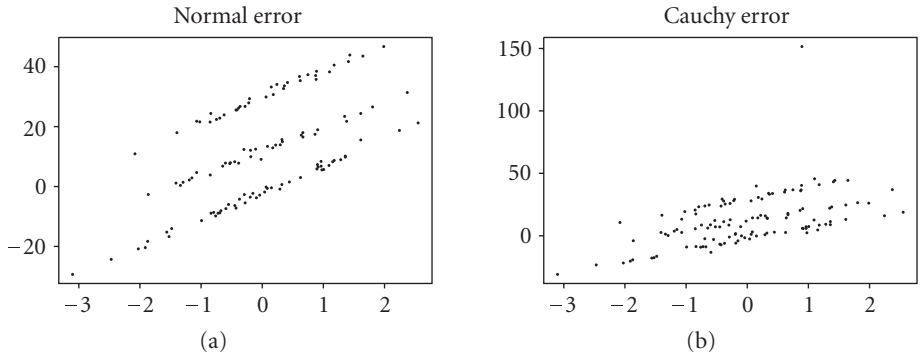

Figure 5.3. Plots of simulated data for three separated linear patterns.

Table 5.2. Relative frequencies of selecting k based on 500 simulations (Case 1).

	$e_j \sim N(0,1)$, N1C1							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C1							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.588	0.772	0.860	0.926
$k=2$	0.160	0.170	0.186	0.216	0.150	0.076	0.042	0.010
$k=3$	0.368	0.392	0.412	0.432	0.130	0.078	0.054	0.042
$k=4$	0.472	0.438	0.402	0.352	0.132	0.074	0.044	0.022

† The true number of clusters $k_0 = 1$.

models while there are some outliers far away from the whole pattern for each case with Cauchy errors.

For each of the aforementioned six models, we generate the data by the model structure (2.1), we then use LS-C and MR-C to select the number of clusters and classify the data. This process is then repeated 500 times separately. To reduce the computation burden, we only fit models with possible numbers of clusters as 1, 2, 3, 4 when the true number of clusters k_0 is 1 or 2; and we only consider possible cluster size of 1, 2, 3, 4, and 5, when k_0 is 3.

In the simulation study, LS-C and MR-C are used to select the best k, respectively. Tables 5.2, 5.3, and 5.4 display the relative frequencies of selecting k for each of the six models using LS-C and MR-C separately. It is apparent that both Huber ρ and LS functions perform extremely well for these models with normal errors. However, as shown in

Table 5.3. Relative frequencies of selecting k based on 500 simulations (Case 2).

	$e_j \sim N(0,1)$, N1C2							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C2							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.086	0.088	0.098	0.114	0.898	0.924	0.956	0.970
$k=3$	0.278	0.294	0.318	0.346	0.054	0.036	0.022	0.018
$k=4$	0.636	0.618	0.584	0.540	0.048	0.040	0.022	0.012

† The true number of clusters $k_0 = 2$.

Table 5.4. Relative frequencies of selecting k based on 500 simulations (Case 3).

	$e_j \sim N(0,1)$, N1C3							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=3$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C3							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.052	0.052	0.052	0.052	0.000	0.000	0.000	0.000
$k=3$	0.148	0.174	0.196	0.214	0.932	0.950	0.958	0.972
$k=4$	0.370	0.382	0.396	0.394	0.056	0.040	0.034	0.022
$k=5$	0.430	0.392	0.356	0.340	0.012	0.010	0.008	0.006

† The true number of clusters $k_0 = 3$.

these tables, in contrast to the nearly perfect performance of both criteria in the normal error models, when the random errors used to generate the data in each case are from Cauchy(0, 1), MR-C with Huber ρ function still selects the underlying true numbers of clusters with promising high proportions of correctness while LS-C loses the power of detecting the underlying numbers of clusters significantly.

Appendix

Let $\mathbb{O}_l$ be any component or a subset of a component of the underlying true partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1^{(n)}, \ldots, \mathbb{O}_{k_0}^{(n)}\}$ of the n objects $\mathbb{O}^{(n)}$. Let $n_l = |\mathbb{O}_l|$. The following lemmas hold in $\mathbb{O}_l$ and can be proved similarly as by Wu and Zen [11].

Lemma A.1. *Suppose that (B1), (C1)-(C2), (X), and (X1) hold. Then,*

$$\frac{1}{n_l} \sum_{j \in \mathbb{O}_l} [\gamma_j - \mathrm{E}(\gamma_j)] \longrightarrow 0, \quad a.s., \tag{A.1}$$

where $\gamma_j = \rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}) - \rho(e_{j,\mathbb{O}_l}) + \mathbf{x}'_{j,\mathbb{O}_l}(\boldsymbol{\beta} - \boldsymbol{\beta}_{0l})\psi(e_{j,\mathbb{O}_l})$ *if* $|\boldsymbol{\beta} - \boldsymbol{\beta}_{0l}| > 0$.

Lemma A.2. *Suppose that the Assumptions (B1)–(B3), (C1)–(C3), (X), and (X1) hold. Then*

$$\sum_{j \in \mathbb{O}_l} \mathbf{x}_{j,\mathbb{O}_l} \psi(e_{j,\mathbb{O}_l}) = O\left((n_l \log\log n_l)^{1/2}\right), \quad a.s. \tag{A.2}$$

$$\sum_{j \in \mathbb{O}_l} [\rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\hat{\boldsymbol{\beta}}_{0l}) - \rho(e_{j,\mathbb{O}_l})] = O(\log\log n_l), \quad a.s. \tag{A.3}$$

$$\hat{\boldsymbol{\beta}}_{0l} = \boldsymbol{\beta}_{0l} + O\left((\log\log n_l / n_l)^{1/2}\right), \quad a.s., \tag{A.4}$$

where

$$\hat{\boldsymbol{\beta}}_{0l} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathbb{O}_l} \rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}). \tag{A.5}$$

Lemma A.3. *Suppose that the Assumptions (B1), (B2), (B3), (C2), (X), and (X1) hold. Then there exists a constant* $\delta > 0$ *such that*

$$\sum_{j \in \mathbb{O}_l} [\rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}^*) - \mathrm{E}(\rho(e_{j,\mathbb{O}_l}))] \geq \delta n_l + o(n_l), \quad a.s. \tag{A.6}$$

holds for all $\boldsymbol{\beta}^* \in \overline{\mathcal{D}}$ *and* n_l *large enough, where* $\overline{\mathcal{D}}$ *is defined in the preceding lemma.*

Acknowledgments

The authors would like to thank the referees for comments and suggestions that improved the presentation of this paper. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] H.-H. Bock, "Probability models and hypotheses testing in partitioning cluster analysis," in *Clustering and Classification*, P. Arabie, L. J. Hubert, and G. De Soete, Eds., pp. 377–453, World Scientific, River Edge, NJ, USA, 1996.

[2] R. E. Quandt and J. B. Ramsey, "Estimating mixtures of normal distributions and switching regressions," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 730–752, 1978.

[3] H. Späth, "A fast algorithm for clusterwise linear regression," *Computing*, vol. 29, no. 2, pp. 175–181, 1982.

[4] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of Classification*, vol. 5, no. 2, pp. 249–282, 1988.

[5] S. Lou, J. Jiang, and K. Keng, "Clustering objects generated by linear regression models," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1356–1362, 1993.

[6] Q. Shao and Y. Wu, "A consistent procedure for determining the number of clusters in regression clustering," *Journal of Statistical Planning and Inference*, vol. 135, no. 2, pp. 461–476, 2005.

[7] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1986.

[8] P. J. Huber, "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, vol. 1, pp. 799–821, 1973.

[9] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, 1996.

[10] J. A. F. Machado, "Robust model selection and M-estimation," *Econometric Theory*, vol. 9, no. 3, pp. 478–493, 1993.

[11] Y. Wu and M. M. Zen, "A strongly consistent information criterion for linear model selection based on M-estimation," *Probability Theory and Related Fields*, vol. 113, no. 4, pp. 599–625, 1999.

[12] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series, no. 28, Princeton University Press, Princeton, NJ, USA, 1970.

[13] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B*, vol. 41, no. 2, pp. 190–195, 1979.

C. R. Rao: Department of Statistics, Penn State University, University Park, PA 16802, USA
Email address: crr1@psu.edu

Y. Wu: Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3
Email address: wuyh@mathstat.yorku.ca

Q. Shao: Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936, USA
Email address: qing.shao@novartis.com