EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 4, No. 4, 2011, 455-466
ISSN 1307-5543 – www.ejpam.com

Estimation and Selection in Regression Clustering

Guoqi Qian[1,*], Yuehua Wu[2]

[1] *Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia*
[2] *Department of Mathematics and Statistics, York University, Toronto, Canada*

Abstract. Regression clustering is an important model-based clustering tool having applications in a variety of disciplines. It discovers and reconstructs the hidden structure for a data set which is a random sample from a population comprising a fixed, but unknown, number of sub-populations, each of which is characterized by a class-specific regression hyperplane. An essential objective, as well as a preliminary step, in most clustering techniques including regression clustering, is to determine the underlying number of clusters in the data. In this paper, we briefly review regression clustering methods and discuss how to determine the underlying number of clusters by using model selection techniques, in particular, the information-based technique. A computing algorithm is developed for estimating the number of clusters and other parameters in regression clustering. Simulation studies are also provided to show the performance of the algorithm.

2000 Mathematics Subject Classifications: 62H30, 68T10, 91C20

Key Words and Phrases: Regression clustering, Least squares estimation, Model selection

1. Introduction

Cluster analysis is an important scientific tool for examining multivariate data with a view to uncovering or discovering clusters or groups of homogeneous observations. It finds clusters in the data such that observations are as "similar" as possible within clusters (internal cohesion or homogeneity), and as "dissimilar" as it could be between clusters (external separation or heterogeneity). Cluster analysis should be distinguished from the related problem of discriminant analysis in that it actually establishes the clusters, whereas in discriminant analysis, known clusterings (or groupings) of some observations are used to categorize others and infer the structure of the data as a whole.

Clustering techniques range from those that are largely heuristic and descriptive to more formal procedures based on statistical models. In general, they follow either a hierarchical strategy or partitioning type of methods. Hierarchical methods proceed by stages producing a series of partitions, which may run from a single cluster containing all objects to as many

[*] Corresponding author.

Email addresses: `g.qian@ms.unimelb.edu.au` (G. Qian), `wuyh@mathstat.yorku.ca` (Y. Wu)

clusters as the total number of objects, with each containing a single object. They can be either "agglomerative", meaning that groups are merged, or "divisive", in which one or more groups are split at each stage.

At each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion. Conventional agglomerative hierarchical methods use heuristic criteria, such as single linkage (nearest neighbor), complete linkage (furthest neighbor), centroid clustering, or sum of squares etc. [11]. In applications, divisive methods are less commonly used than agglomerative procedures since they are computationally demanding.

Yet a significant drawback of hierarchical clustering methods is that the divisions or fusions, once made, are irrevocable. When an agglomerative algorithm has joined two objects into a cluster they cannot subsequently be separated, and when a divisive algorithm has made a split, the objects cannot be recombined. As Kaufman and Rousseeuw [11] comment: "A hierarchical method suffers from the defect that it can never repair what was done in previous steps".

In contrast, a partitioning method constructs a fixed number of clusters, say k. It classifies the data into k clusters, which together satisfy two requirements of a partition: (i) each cluster must contain at least one object; (ii) each object must belong to exactly one cluster. Usually, partitioning methods move observations iteratively from one cluster to another, starting from an initial partition, to achieve some pre-chosen optimization. In most circumstances, the number of clusters has to be specified in advance and typically does not change during the course of the iteration. For instance, the most commonly used relocation methods – the k-means type of methods: k-means, k-modes, k-medians and k-mediods [9, 10] – reduce the average within-group distance of objects to their nearest representatives (means, modes, medians or medoids).

We can easily envision that to identify possible clusters of observations in data, it is of essential importance to have the knowledge of how "close" individuals are to each other, or how far apart they are. The aforementioned methods such as the single linkage, complete linkage, k-means, etc. are usually considered as descriptive methods since they are mainly heuristically motivated and use descriptive statistics as the measures of similarity or dissimilarity between observations. For instance, the k-means type of methods are characterized by taking the distance (Euclidean or Manhattan or Minkowski distances) of each object to the cluster centres (mean, median, mode, medoid) as the similarity or dissimilarity measure. On the other hand, model-based clustering uses a probability model as the similarity or dissimilarity measure, i.e. objects have the same model specification within clusters. Furthermore, model-based clustering techniques use inferential statistics by means of probabilistic models, not only for checking the significance of clusters and clustering, but also for providing a firm theoretical basis for clustering methods and strategies. Model-based methods can be applied both in hierarchical clustering and partitioning-type clustering. This research focuses on partitioning-type model-based approaches.

It is noted that there is no absolute boundary between descriptive and model-based clustering methods. Some clustering methods were heuristically motivated, but later on statisticians studied their performance from a probabilistic perspective. For instance, [10, 3] studied the asymptotic behaviour of k-means using a model-based approach; [8, 12] investigated the mathematical relationship between high-density clusters and the single-linkage clustering method.

Consider a finite set of n objects $\mathcal{O} = \{1, \ldots, n\}$ together with data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^p$ describing the properties of these objects. Based on these data, our problem is to recover the latent partitioning $\Pi = (\mathscr{C}_1, \ldots, \mathscr{C}_k)$ of $\mathcal{O}$ and to construct a clustering of the corresponding objects. A model-based or probabilistic clustering approach assumes that the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are a sample of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ that belong to a structured population. It characterizes the clusters by specifications for the probability distribution of the random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ which may differ from cluster to cluster.

Roughly speaking, stochastic model-based clustering techniques can be divided into the following two categories: (1) Parametric approach in which the probability distribution of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ is assumed to have a known parametric form but with unknown parameters; (2) Non-parametric approach in which no distributional assumption is explicitly made for the individual clusters.

This paper will focus on parametric model-based partitioning-type clustering methods, in particular, the likelihood method. Further, we study only the regression clustering problem, which means the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ contain observations of both dependent and explanatory variables and their relationship is of our interest. In Section 2, we review regression clustering. In Section 3, we discuss a procedures for estimating the parameters and the number of clusters in linear regression clustering under the classification likelihood framework. In Section 4, an algorithm is given for selecting the clustering and the number of clusters in regression clustering. The simulation study is presented in Section 5. Finally, Section 6 provides some discussions and concludes the paper.

2. Regression Clustering

Regression clustering refers to estimating the class-specific regression hyperplanes underlying the data that randomly come from a population consisting of distinct classes. Note that the notion hyperplane used here is a generic one, which means it does not necessarily pass through the origin in the space. It should be more correctly called an affine set. But we do not distinguish them in this paper.

For the regression clustering problem, the data have the form $(y_j, \boldsymbol{x}'_j), j = 1, \ldots, n$, where $\boldsymbol{x}_j \in \mathbb{R}^p$ is a (non-random) explanatory column vector and $y_j \in \mathbb{R}$ a random dependent variable for the j-th object. As in the general setting of model-based clustering, there are also two different approaches for regression clustering in the literature. One is the random partition regression clustering. The discussion can be found in [16, 15] among others. Another one is the fixed partition regression clustering. As discussed in [4, 5, 6, 7], the classification likelihood model or the fixed partition regression clustering model for any partition $\Pi = (\mathscr{C}_1, \ldots, \mathscr{C}_k)$ of $\mathcal{O}$ is:

$$Y_j \sim f(\cdot; \boldsymbol{\beta}_i, \sigma_i) \sim \phi(\boldsymbol{x}'_j \boldsymbol{\beta}_i, \sigma_i) \text{ for all } j \in \mathscr{C}_i,\ i = 1, \ldots, k.$$

Equivalently, it can be written in the form of a group of linear models:

$$y_j = \boldsymbol{x}'_j \boldsymbol{\beta}_i + e_j, \quad e_j \sim N(0, \sigma_i^2) \text{ for all } j \in \mathscr{C}_i,\ i = 1, \ldots, k. \tag{1}$$

Under the fixed-partition model (1), the log-likelihood function is given by

$$\log L_n(k,(\boldsymbol{\beta}_i,\sigma_i^2)_{i=1,\dots,k}) = -\frac{1}{2}\sum_{i=1}^{k}\sum_{j\in\mathscr{C}_i}\left(\log 2\pi + \log\sigma_i^2 + \frac{(y_j-\boldsymbol{\beta}_i'\boldsymbol{x}_j)^2}{\sigma_i^2}\right). \tag{2}$$

For given $(\hat{\boldsymbol{\beta}}_i,\hat{\sigma}_i^2)_{i=1,\dots,k}$, (2) is maximized at setting $\mathscr{C}_i$ to

$$\hat{\mathscr{C}}_i = \arg\min_i \left(\log\hat{\sigma}_i^2 + \frac{(y_j-\hat{\boldsymbol{\beta}}_i'\boldsymbol{x}_j)^2}{\hat{\sigma}_i^2}\right). \tag{3}$$

For given $\hat{\mathscr{C}}_i$, (2) is the sum of the usual log-likelihood functions for homogeneous linear regressions within clusters. Hence, it is maximized by the LS-estimator $\hat{\boldsymbol{\beta}}_i$ from the data points $(y_j,\boldsymbol{x}_j)$ with $j\in\mathscr{C}_i$ and

$$\hat{\sigma}_i^2 = \frac{\sum_{j\in\hat{\mathscr{C}}_i}(y_j-\hat{\boldsymbol{\beta}}_i'\boldsymbol{x}_j)^2}{\hat{n}_i}, \quad i=1,\dots,k, \tag{4}$$

where $\hat{n}_i = |\hat{\mathscr{C}}_i|$ is the number of data points in $\mathscr{C}_i$. Then $\log\hat{L}_n$ is monotonically increased if the steps (3) and (4) are carried out alternately. This algorithm leads to a local maximum (one would hope, to an approximation of the global maximum by proper initialization) in finitely many steps.

The fixed partition approach has a particular advantage over the random partitioning in the context of regression clustering. As observed by Hennig [1], the mixture model presumes implicitly an *assignment independence* of each object to clusters with respect to the covariate vectors $\boldsymbol{x}_j$. That is, the clusters keep the same conditional proportions $\pi_i, i=1,\dots,k$ for every fixed covariate vector $\boldsymbol{x}_j$. In other words, the probability of a point $(y_j,\boldsymbol{x}_j')$ to be generated by cluster i is independent of $\boldsymbol{x}$ and j. This is generally not true as shown in Figure 1, which is adapted from [1]. On the other hand, the fixed partition model (1) supposes that the cluster membership of each object or cluster labels are explicitly parametrized and are determined by the estimation of $\widehat{\boldsymbol{\beta}}_i$ and $\hat{\sigma}_i^2$ through the points $(y_j,\boldsymbol{x}_j')(j\in\mathscr{C}_i)$. Hence the fixed partition model does make allowance of possible *assignment dependence* between the j-th object and the associated covariate $\boldsymbol{x}_j$.

3. Procedures for Estimating the Parameters and the Number of Clusters in Regression Clustering

Suppose that we have n objects $\mathscr{O}^{(n)}=\{1,2,\dots,n\}$ with the associated data points $(\boldsymbol{x}_1,y_1)$, …, $(\boldsymbol{x}_n,y_n)$. Here $\boldsymbol{x}_j\in\mathbb{R}^p$ is a non-random explanatory p-vector and $y_j\in\mathbb{R}$ is a random dependent variable for the j-th object ($j=1,2,\dots,n$). These n objects are assumed to be a random sample coming from a structured population, which consists of a fixed (but unknown) number, say k_0, of sub-populations each of which is characterized by a regression hyperplane with class-specific unknown parameters. Therefore for the n observations from

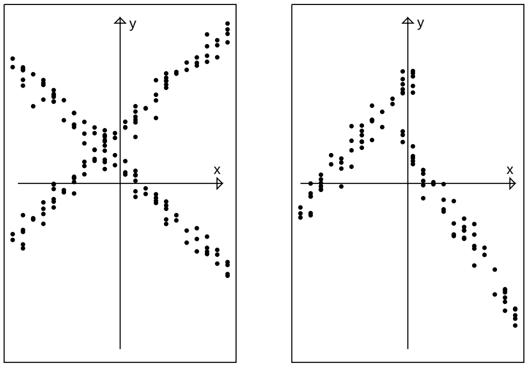


Figure 1: Assignment independence – assignment dependence

this population, there exists an underlying partition $\Pi_{k_0}^{(n)} = \{\mathscr{O}_1^{(n)}, \ldots, \mathscr{O}_{k_0}^{(n)}\}$, and each cluster $\mathscr{O}_i^{(n)} \triangleq \{i_1, \ldots, i_{n_i}\} \subseteq \mathscr{O}^{(n)}$ is represented by

$$\boldsymbol{y}_{\mathscr{O}_i} = X_{\mathscr{O}_i}\boldsymbol{\beta}_{0i} + \boldsymbol{e}_{\mathscr{O}_i}, \quad \boldsymbol{e}_{\mathscr{O}_i} \sim N(0, \sigma_i^2 I_{n_i}), \tag{5}$$

where $\boldsymbol{y}_{\mathscr{O}_i} = (y_{i_1}, \ldots, y_{i_{n_i}})'$, $X_{\mathscr{O}_i} = (\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_{n_i}})'$ is an $n_i \times p$ design matrix in the cluster $\mathscr{O}_i$, $\boldsymbol{e}_{\mathscr{O}_i}$ is an n_i-vector of random errors, I_{n_i} is an $n_i \times n_i$ identity matrix, and $n_i = |\mathscr{O}_i|$ for $i = 1, \ldots, k_0$. Here, $(\boldsymbol{\beta}_{0i}', \sigma_i)' \in \mathbb{R}^p \times \mathbb{R}^+$, $1 \leq i \leq k_0$, are k_0 unknown parameter vectors. $\boldsymbol{\beta}_{0i}$, $1 \leq i \leq k_0$, are assumed to be distinct from one another. It is clear that $n = n_1 + \ldots + n_{k_0}$. In the following, we assume that $k_0 \leq K$, where K is a known positive integer. Note that in (5) we have suppressed the n in $\mathscr{O}_i^{(n)}$ for convenience.

The objective is to reconstruct the underlying structure (5) from the observed data by estimating the number of clusters k_0 and then classifying the data and estimating the class-specific parameters accordingly. What can be done in practice, however, is to first consider every given partition of these n observations: $\Pi_k^{(n)} = \{\mathscr{C}_1^{(n)}, \ldots, \mathscr{C}_k^{(n)}\}$, where $k \leq K$ is a positive integer. For such a partition, one then fits k clusterwise regression models and obtain k Least Squares (LS) estimates $\widehat{\boldsymbol{\beta}}_i$, $i = 1, \ldots, k$. By this stage, one can use a criterion to select the best k and the associated partition. Shao and Wu [14] propose an information-based criterion for determining the number of clusters as following: Let $q(k)$ be a strictly increasing positive function of k, and A_n be a sequence of positive constants. Define

$$D_n(\Pi_k^{(n)}) = \sum_{i=1}^{k} ||\boldsymbol{y}_{\mathscr{C}_i^{(n)}} - X_{\mathscr{C}_i^{(n)}}\hat{\boldsymbol{\beta}}_i||^2 + q(k)A_n\,, \tag{6}$$

and $\hat{k}_n$, the estimate of k_0, is the integer that minimizes this criterion, i.e.

$$D_n(\hat{k}_n) = \min_{1 \leq k \leq K} \min_{\Pi_k^{(n)}} D_n(\Pi_k^{(n)}), \tag{7}$$

where $1 \leq k \leq K$. It can be seen that in (6), the first term is the residual sum of squares which measures the goodness of fit of the model and the second term is the penalty for over-fitting. Furthermore, the criterion (7) shows that one determines the optimal number of clusters and the corresponding partitioning simultaneously. We shall call (7) Criterion LS-C in the sequel, which stands for clustering by the LS method.

Under some mild conditions, it is shown in Shao and Wu [14] that the proposed criterion selects the true number of regression hyperplanes with probability one among all class-growing sequences of classifications, when the number of observations n from the population increases to infinity.

Note that the assumption $\boldsymbol{e}_{\mathscr{O}_i} \sim N(0, \sigma_i^2 I_{n_i})$ in (5) is not required in computing the LS-estimates $\widehat{\boldsymbol{\beta}}_i$ and the criterion function $D_n(\Pi_k^{(n)})$. But the least squares estimates are known to be sensitive to outliers and violation of the normality assumption in the data. This implies that the LS-C criterion is expected to work well for selecting the number of clusters and estimating the partition in linear regression clustering only when the normality assumption is not seriously violated. Recently, a consistent robust procedure for determining the number of clusters in regression clustering is proposed in [2]. However, we will not get into its theoretic detail here to keep this paper into reasonable length. We will use only the simulation in section 5 to illustrate the sensitivity of the LS-C criterion against normality.

Finally, it seems that each squared residual sum in the first term of (6) should be scaled by the corresponding variance estimate $\hat{\sigma}_i^2$. Actually ignoring this scaling does not affect the asymptotic properties of the criterion function $D_n(\Pi_k^{(n)})$. It turns out that ignoring the scaling would improve the robustness of the clustering procedure. This is because a large σ_i^2 estimate is more likely to be associated with a cluster with large variability, thus being less separable from the other clusters. Ignoring the scaling would favor not including the outlaying data points in the current cluster.

4. An Algorithm for Estimation and Selection in Regression Clustering

We give an iterative algorithm in this section to implement the procedures in the previous section for selecting the optimal clustering and estimating the number of clusters in regression clustering.

For each fixed k, we obtain the optimal clustering of the data $\Pi_k = \{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$ by minimizing the within-cluster sum of residual squares. The quantity to be minimized is then

$$\mathrm{SRSS}(\Pi_k) = \sum_{i=1}^{k} ||\boldsymbol{y}_{\mathscr{C}_i} - X'_{\mathscr{C}_i}\widehat{\boldsymbol{\beta}}_i||^2 \tag{8}$$

where $\widehat{\boldsymbol{\beta}}_i$, $i = 1, \cdots, k$, are the least squares estimators based on given $\{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$. This minimization can be accomplished according to the following algorithm:

(i) Label all the data points in the sample as 1 to n. Given an initial partition
$\Pi_k = \{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$ of $\mathscr{O} = \{1, \ldots, n\}$, fit regression models for each of the k clusters and obtain the overall sum of the squared residuals SRSS_0 for this partition. Initialize $i = 0$.

(ii) Set $i = i + 1$ and reset $i = 1$ if $i > n$. Suppose $i \in \mathscr{C}_j$. Then move i into $\mathscr{C}_h$, $h = 1, \ldots, k$, $h \neq j$ respectively. For each of these $k - 1$ relocations, re-fit the regression models for the changed clusters and calculate the overall sum of the squared residuals accordingly. Denote the smallest one by SRSS_h. If $\text{SRSS}_h < \text{SRSS}_0$, redefine $\mathscr{C}_j = \mathscr{C}_j - \{i\}$, $\mathscr{C}_h = \mathscr{C}_h + \{i\}$, and set $\text{SRSS}_0 = \text{SRSS}_h$. Otherwise keep i in $\mathscr{C}_j$.

(iii) Repeat (ii) until the objective function (8) could not be reduced any further, which means no observation relocation is necessary and the optimal clustering is achieved for this k.

The idea behind the above algorithm comes from [7]. Once the optimal clustering is done for each possible k, the Criterion LS-C is used as a rule to select the best number of clusters.

It is noted that the initial partition of $\mathscr{O} = \{1, \ldots, n\}$, if properly set, will facilitate the convergence and performance of the algorithm above. Denote the complement set of a set C by C^c. We propose to generate an initial partition of a dataset as follows:

Step 1. Consider the linear model

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + e_i. \tag{9}$$

Based on the whole dataset, one estimates $\boldsymbol{\beta}$ by a robust method, e.g. least median squares method or least trimmed squares method [13].

Step 2. Put all data points, whose distances to the regression hyperplane estimated in Step 1 is less than a predetermined number, say δ, into a set C_1. If $|C_1|$ and $|C_1^c|$ are both larger than a predetermined integer, say m, set $\ell = 1$ and go to the next step; otherwise, set $\ell = 0$ and go to Step 5.

Step 3. Based on the dataset $\bigcap_{i=1}^{\ell} C_i^c$, one estimates $\boldsymbol{\beta}$ in (9) by the same robust method used in Step 1.

Step 4. Put all data points in $\bigcap_{i=1}^{\ell} C_i^c$, whose distances to the regression hyperplane estimated in Step 3 is less than δ, into a set $C_{\ell+1}$. If $|C_{\ell+1}|$ and $|\bigcap_{i=1}^{\ell+1} C_i^c|$ are both larger than m, set $\ell = \ell + 1$ and repeat Step 3; otherwise, go to Step 5.

Step 5. The initial partition is $\{C_1, \ldots, C_\ell, \bigcap_{i=1}^{\ell} C_i^c\}$ if $\ell > 1$ or just the whole dataset itself if $\ell = 0$.

5. Simulation study

In this section we assess the finite sample performance of Criterion LS-C together with the use of the algorithm in the previous section. Simulated data sets are to be used to perform regression clustering for the assessment.

While many types of data sets can be simulated, we consider only two factors in determining the type: number of clusters (2 or 3), and error distributions (standard normal $N(0,1)$ or $t(3)$), so there are in total 4 cases of data to be considered, which are summarized in Table 1.

There will be only one covariate involved in the regression in each cluster, and the covariate is generated from $N(0,1)$. The parameters used for each case are given in Table 2. Then the fixed partition regression clustering model $y_{ji} = \boldsymbol{x}'_{ji}\boldsymbol{\beta}_{0i} + e_{ji}, j = 1,\ldots,n_i,\ i = 1,\ldots,k_0$ is applied to generate the response values y_{ji}, where e_{ji} is a random number originating from $N(0,1)$ or $t(3)$, and the first element of $\boldsymbol{x}_{ji}$ is the constant 1 corresponding to the intercept term in the model.

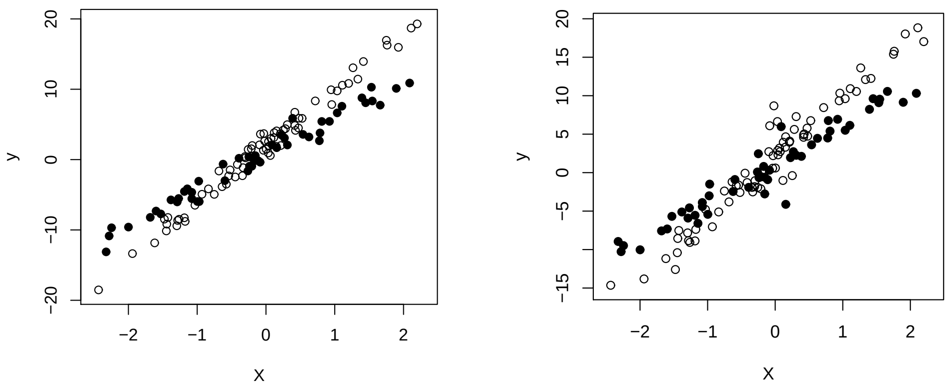
Table 1: Shorthand notation for the four cases.

N1C2	Case 1,	two regression lines	Normal error
T1C2	Case 2,	two regression lines	$t(3)$ error
N1C3	Case 3,	three regression lines	Normal error
T1C3	Case 4,	three regression lines	$t(3)$ error

Table 2: Parameter values used in the simulation study of regression clustering.

Case	k_0	Regression coefficients	No. of obs.
1–2	2	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 2 \\ 8 \end{pmatrix}, \boldsymbol{\beta}_{02} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}$	$n_1 = 70$ $n_2 = 50$
3–4	3	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 18 \\ 6 \end{pmatrix}, \boldsymbol{\beta}_{02} = \begin{pmatrix} 12 \\ 8 \end{pmatrix}, \boldsymbol{\beta}_{03} = \begin{pmatrix} 15 \\ -2 \end{pmatrix}$	$n_1 = 35$ $n_2 = 35$ $n_3 = 50$

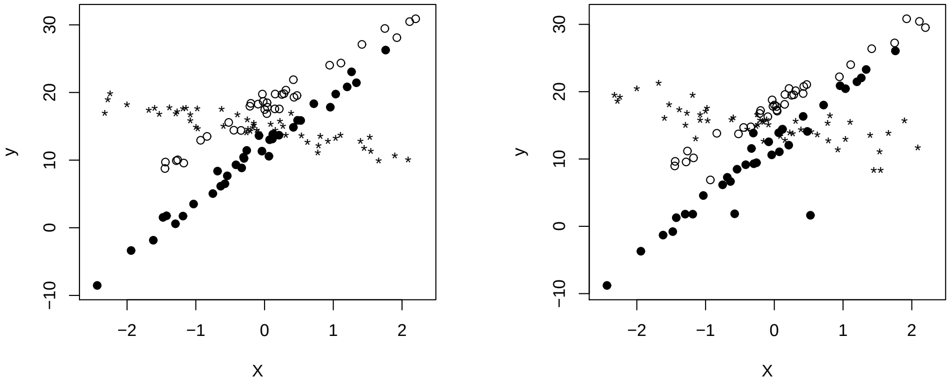
Figures 2 and 3 illustrate what the data typically would look like for Cases 1 to 4 with Normal or $t(3)$ errors. These figures show that the groupings of the linear patterns are visible with standard normal random errors and getting worse with $t(3)$ random errors.



(a) Case 1

(b) Case 2

Figure 2: Simulated data with two clusters.



(a) Case 3

(b) Case 4

Figure 3: Simulated data with three clusters.

In this study, we set $q(k) = kp$ in Criterion LS-C (6), where p is the number of regression coefficients in the model and is a constant in our study; and k is the number of clusters used in the regression clustering under assessment. It is noted that in an information model selection criterion, a penalty function, which is A_n in (6), is usually chosen as $c\log(n)$ or $c\log\log(n)$ with a constant $c > 0$. In light of the fact that $\lim_{\lambda\to 0}\left[(\log n)^{\lambda} - 1\right]/\lambda = \log\log n$, we set $A_n = \left[(\log n)^3 - 1\right]/3$.

For each of the four cases, we conduct 1000 simulations using Criteria LS-C. To apply the algorithm given in Section 4, we set $\delta = 0.2$ and $m = 2p$. The algorithm is then used to estimate the number of clusters in linear regression clustering. In Table 3 we summarize the results from the simulation study, where each number represents the relative frequency of selecting a given number k clusters in regression clustering out of the 1000 replications.

From Table 3 we see that Criterion LS-C performs almost perfectly in Cases 1 and 3, which is expected since the errors are standard normal distributed. However, the criterion tends to over-estimate the number of clusters when the error distribution becomes heavy-tailed, as shown in Cases 2 and 4. This is also expected but it indicates that the direction of non-robustness of LS-C against normality is more likely to be over-clustering rather than under-clustering.

The cluster-specific regression lines can also be estimated during applying the criterion LS-C. Table 4 presents the estimation of the regression parameters by applying LS-C to the data shown in Figures 2 and 3. From the table, one can conclude that when the errors are $t(3)$ distributed, the least squares regression clustering method is not able to capture the underlying groupings, while it can when the errors are normal.

Table 3: Relative frequencies of selecting k based on 1000 simulations for Cases 1-4.

	Case 1	Case 2	Case 3	Case 4
	$N(0,1)$ error	$t(3)$ error	$N(0,1)$ error	$t(3)$ error
	$k_0 = 2$	$k_0 = 2$	$k_0 = 3$	$k_0 = 3$
$k = 1$	.000	.001	.000	.000
$k = 2$	.986	.422	.000	.000
$k = 3$	.014	.488	.999	.791
$k = 4$	.000	.087	.001	.207
$k = 5$	.000	.002	.000	.002

Table 4: The estimation of the regression parameters by applying LS-C to the data shown in Figures 2 and 3

k_0	Case	Clusters	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	β_3	β_4
2		True	$\begin{pmatrix} 2 \\ 8 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 5 \end{pmatrix}$		
	1	LS-C	$\begin{pmatrix} 2.12 \\ 8.02 \end{pmatrix}$	$\begin{pmatrix} 0.76 \\ 5.11 \end{pmatrix}$		
	2	LS-C	$\begin{pmatrix} 1.48 \\ 5.56 \end{pmatrix}$	$\begin{pmatrix} -1.13 \\ 5.87 \end{pmatrix}$	$\begin{pmatrix} 4.46 \\ 6.18 \end{pmatrix}$	
3		True	$\begin{pmatrix} 18 \\ 6 \end{pmatrix}$	$\begin{pmatrix} 12 \\ 8 \end{pmatrix}$	$\begin{pmatrix} 15 \\ -2 \end{pmatrix}$	
	3	LS-C	$\begin{pmatrix} 18.05 \\ 6.06 \end{pmatrix}$	$\begin{pmatrix} 11.97 \\ 8.02 \end{pmatrix}$	$\begin{pmatrix} 14.66 \\ -1.85 \end{pmatrix}$	
	4	LS-C	$\begin{pmatrix} 17.74 \\ 6.14 \end{pmatrix}$	$\begin{pmatrix} 12.02 \\ 8.16 \end{pmatrix}$	$\begin{pmatrix} 10.73 \\ -2.87 \end{pmatrix}$	$\begin{pmatrix} 15.54 \\ -1.70 \end{pmatrix}$

6. Discussion

In this paper we review the general cluster analysis methods, then focus on regression clustering which uses the model-based fixed partition method and also takes into account the dependence between the response and explanatory variables. Regression clustering has not been widely used in practice even though it has a great potential. A possible reason is the computing complexity involved in the method. This paper provides a computing procedure and a feasible algorithm for estimation and selection involved in regression clustering. The simulation study concludes a satisfactory finite sample performance of the algorithm when the error distribution involved is close to normal. It also suggests the need to use a robust clustering method when the error distribution strays away from the normal.

References

[1] C Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296, 2000.

[2] C Rao and Y Wu and Q Shao. An M-Estimation-Based Procedure for Determining the Number of Regression Models in Regression Clustering. *Journal of Applied Mathematics and Decision Sciences*, 2007, 2007.

[3] D Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9:135–140, 1981.

[4] H Bock. The equivalence of two extremal problems and its application to the iterative classification of multivariate data. Manuscript for the medizinische statistik conference, Forschungsinstitut Oberworfachl, 1969.

[5] H Bock. Probability models and hypotheses testing in partitioning cluster analysis. In P Arabie and L Hubert and G De Soete, editor, *Clustering and Classification.*, pages 377–453, River Edge, New Jersey., 1996. World Scientific Publishing.

[6] H Späth. Clusterwise linear regression. *Computing*, 22:367–373, 1979.

[7] H Späth. Algorithm 48: A fast algorithm for clusterwise linear regression. *Computing*, 29:175–181, 1982.

[8] J Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76:388–394, 1981.

[9] J Hartigan and M Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1978.

[10] J MacQueen. Some methods for classification and analysis of multivariate observations. In N Le Cam and J Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.*, volume 1, pages 281–297. University of California Press., 1967.

[11] L Kaufman and P Rousseeuw. *Finding Groups in Data*. Wiley-Interscience, New York, 1990.

[12] M Wong. A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association*, 77:841–847, 1982.

[13] P Rousseeuw and A Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[14] Q Shao and Y Wu. A consistent procedure for determining the number of clusters in regression clustering. *Journal of Statistical Planning and Inference*, 135:461–476, 2005.

[15] R Quandt and J Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association.*, 73:730–752, 1978.

[16] W DeSarbo and W Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5:249–282, 1988.