Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 4037380, 13 pages
http://dx.doi.org/10.1155/2016/4037380

Research Article

Semisupervised Clustering by Iterative Partition and Regression with Neuroscience Applications

Guoqi Qian,[1] Yuehua Wu,[2] Davide Ferrari,[1] Puxue Qiao,[1] and Frédéric Hollande[3]

[1]*School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia*
[2]*Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3*
[3]*Department of Pathology, University of Melbourne, Parkville, VIC 3010, Australia*

Correspondence should be addressed to Yuehua Wu; wuyh@mathstat.yorku.ca

Received 16 December 2015; Accepted 17 March 2016

Academic Editor: Guoli Ji

Regression clustering is a mixture of unsupervised and supervised statistical learning and data mining method which is found in a wide range of applications including artificial intelligence and neuroscience. It performs unsupervised learning when it clusters the data according to their respective unobserved regression hyperplanes. The method also performs supervised learning when it fits regression hyperplanes to the corresponding data clusters. Applying regression clustering in practice requires means of determining the underlying number of clusters in the data, finding the cluster label of each data point, and estimating the regression coefficients of the model. In this paper, we review the estimation and selection issues in regression clustering with regard to the least squares and robust statistical methods. We also provide a model selection based technique to determine the number of regression clusters underlying the data. We further develop a computing procedure for regression clustering estimation and selection. Finally, simulation studies are presented for assessing the procedure, together with analyzing a real data set on RGB cell marking in neuroscience to illustrate and interpret the method.

1. Introduction

Regression and clustering are probably two of the most important statistical data mining methods used in practice including artificial intelligence and neuroscience. However, regression clustering, a data mining method integrating the two, has rarely been studied as a single entity despite its great potential for practical use. It is, thus, the intention of this paper to focus on statistical estimation, selection, and computing of regression clustering. In this section, we briefly review cluster analysis but not the familiar regression analysis and then introduce the regression clustering problem.

(1) Cluster Analysis. Cluster analysis is an important unsupervised statistical learning and data mining technique for clustering homogeneous observations from data. Its main objective is to divide a collection of data points, often of multivariate nature, into subsets or "clusters" such that observations within one cluster are more "similar" (homogeneous) to each other than to observations in different clusters. Cluster analysis is usually used in situations where clustering information is not observed on the data points and one wants to get this information from the data to explicitly group them.

Many approaches have been developed in cluster analysis, which in general fall into two categories: hierarchical and partitive. A hierarchical approach proceeds by either a sequence of "agglomerative" stages or a sequence of "divisive" ones. At each agglomerative stage, clusters are produced by merging or retaining the clusters produced at the immediate previous stage, where clusters at the initial stage may simply be taken to be those individual data points. Contrarily, at each divisive stage, clusters are produced by splitting or retaining the clusters produced at the immediate previous stage, where one may assume a single cluster containing all the data points at the initial stage. The key feature of a hierarchical approach is that clusters obtained at one stage are derived from those in the immediate previous stage. On the other hand, partitive approaches refer to those nonhierarchical ones which may be further classified according to other features of clustering.

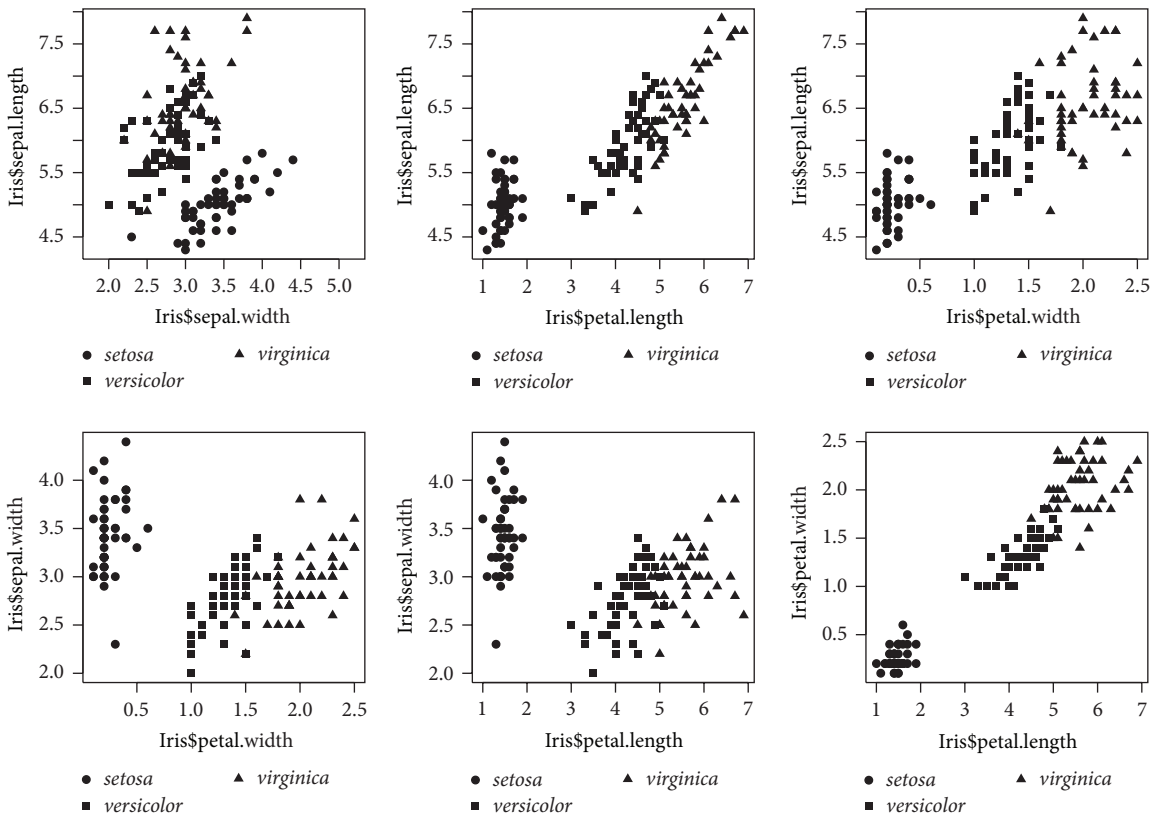


Figure 1: Pairwise scatterplots for the *Iris* data.

The outcome of a hierarchical clustering is often represented by a graph called dendrogram in which each stage of merging or splitting is determined by optimizing some similarity or dissimilarity criterion. A significant drawback of hierarchical clustering methods is that the divisions or fusions, once made, are irrevocable. That is, when an agglomerative algorithm has joined two objects into one cluster, they cannot subsequently be separated, and when a divisive algorithm has made an unwanted split, the objects involved can no longer be recombined into one cluster. Kaufman and Rousseeuw [1] comment on this as follows: "A hierarchical method suffers from the defect that it can never repair what was done in previous steps."

In contrast, a partitive clustering constructs a fixed number of clusters often by an iterative procedure. It imposes two requirements in the procedure: (i) each cluster must contain at least one object and (ii) each object must belong to exactly one cluster. In addition, the number of clusters constructed stays fixed during the iterations and an initial partition is required to start the iteration. At each iteration, a tentative partition is constructed by relocating the data points to optimize a conditional criterion. This procedure continues until certain convergence or stability of partition occurs. Commonly used partitive clustering approaches include those k-means type of methods: k-means, k-modes, k-medians, and k-medoids [2, 3]. New developments in this regard can be found in Hastie et al. [4, section 14.3] and Clarke et al. [5, Chapter 8], for example.

Table 1: Confusion matrix between k-means clustering and species information.

	setosa	*versicolor*	*virginica*
Cluster 1	50	0	0
Cluster 2	0	48	14
Cluster 3	0	2	36

We present an example here to illustrate the use of k-means method for clustering. The example uses the well-known *Iris* data from Anderson [6] which was analyzed by Fisher [7] and many others. The data give the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of the 3 species of *Iris*: *setosa*, *versicolor*, and *virginica*. The data which can be retrieved from statistics package R [8] are displayed in Figure 1, where we see the data of sepal length and width and petal length and width distributed in clusters. So we use the k-means algorithm of Hartigan and Wong [3] to find a partition of 3 clusters for the data and compare the partition with the species information given. The computing is done in R with a random initial partition determined by `set.seed(123)`. The result is summarized in Table 1, from which we see a perfect match between cluster 1 and species

setosa and some mismatch between clusters 2 and 3 and species *versicolor* and *virginica*.

(2) Regression Clustering. In this paper, we will focus on regression clustering, a data mining method which iteratively clusters data into clusters according to the available regression pattern and then updates the regression in each cluster simultaneously until equilibrium is attained. It is commonly known that regression is for studying the relationship between a dependent variable and a set of explanatory variables which have observations on a sample of objects. If the samples come from different populations and the variable indexing the populations also has an effect on the dependent variable, the regression should be performed on individual populations separately through the corresponding subsamples observed, or by including the population effect in the model, in order to make valid or more reliable statistical inference. However, the population indexing variable sometimes is not observed or unobservable. In such situations, it is necessary to cluster the sample objects to conform to their respective populations as much as possible and then apply regression to each cluster. We refer to this procedure as regression clustering if our focus is clustering the data points or as cluster regression if it is studying the unobserved regression patterns in the data.

Before getting into details of regression clustering, we review various measures of similarity or dissimilarity used in general cluster analysis. Note that to identify possible clusters of observations in data it is essential to be able to measure how close or how far individual data objects are to/from each other. Current measures include the single linkage (nearest neighbour) and complete linkage (further neighbour) (cf. [1]) and the k-means. These are usually considered as descriptive since they do not involve any probability distribution and use only descriptive statistics as the measures of similarity or dissimilarity between observations. An obvious disadvantage of using a descriptive measure is one cannot make statistical inference on results of clustering; thus, one is not able to assess the variability involved in the results. To enable making statistical inference, probability distributions or models are postulated for the clusters of data, and it is deemed that data in the same cluster have the same probability distribution. Hence, the similarity or dissimilarity measures to be used are assigned a probability distribution, and the significance and variability of clustering can be readily derived. Probability model based approaches can be applied in both hierarchical and partitive types of clustering. We choose to use the probability model for partitive regression clustering here.

Note that there is no absolute boundary between descriptive and probability model based clustering methods. Some clustering methods were heuristically motivated, but later on, statisticians studied their performance from a probabilistic perspective. For instance, MacQueen [2] and Pollard [9] studied the asymptotic behaviour of k-means using a probability model based approach; Hartigan [10] and Wong [11] 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 $\mathbf{z}_1 = (y_1, \mathbf{x}_1')', \ldots, \mathbf{z}_n = (y_n, \mathbf{x}_n')' \in \mathbb{R}^{p+1}$ being the observations of these objects. The problem of regression clustering is to recover the latent partitioning $\Pi = (\mathscr{C}_1, \ldots, \mathscr{C}_k)$ of $\mathcal{O}$ so that the relationship between $(y_1, \ldots, y_n)'$ and $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ can be studied by regressions on $\mathscr{C}_1, \ldots, \mathscr{C}_k$ separately. A probability model based clustering approach assumes that the observed data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are a sample of respective random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ that belong to a set of populations indexed by Π. Thus, those $\mathbf{Z}_j$ with $j \in \mathscr{C}_i$, $i = 1, \ldots, k$, have the same probability distribution. The specification of the probability distributions can be either parametric or nonparametric.

(3) Organization of the Paper. In Section 2 we provide a detailed formulation of regression clustering including modeling, parameter estimation, and partition determination. In Section 3 we present two procedures for estimating the number of clusters in cluster linear regression. In Section 4, a pointwise iterative assessing algorithm is developed for implementing the regression clustering procedures. A simulation study and an example are presented in Section 5. A real data example on RGB cell marking clustering is analyzed in Section 6. The paper ends with a Conclusion section.

2. Regression Clustering Model and Optimization

Regression clustering becomes very useful when one intends to recover or estimate the unobserved class-specific regression hyperplanes based on the sample data of dependent and explanatory variables. Note that the notion of 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 cluster regression or regression clustering problem, the data have the form $(y_j, \mathbf{x}_j')$, $j = 1, \ldots, n$, where $\mathbf{x}_j \in \mathbb{R}^p$ is an explanatory column vector and $y_j \in \mathbb{R}$ is a random dependent variable for the jth object. The probability distribution of $\mathbf{x}_j$ does not provide any information on regression hyperplanes; thus, our statistical inference will be made conditional on the observed $\mathbf{x}_j$. In other words, we can simply treat $\mathbf{x}_j$ as nonrandom. As in the general setting of probability model based cluster analysis, there are two different approaches for regression clustering. One is the random partition or soft partition approach in which each data point is assigned a nonzero probability to fall into any of the clusters or equivalently follows a mixture probability distribution. The discussion can be found in DeSarbo and Cron [12] and Quandt and Ramsey [13], among others. Another one is the fixed partition or hard partition approach in which each data point is assigned a cluster membership or label through certain optimization procedure, so a data point belongs to only one cluster. As discussed in Bock [14, 15] and Späth [16, 17], the probability distribution or classification likelihood function of a data point in a fixed partition approach of regression clustering, with an unknown partition $\Pi = (\mathscr{C}_1, \ldots, \mathscr{C}_k)$ of $\mathcal{O}$, is of the form

$$Y_j \sim f\left(\cdot; \boldsymbol{\beta}_i, \sigma_i\right) \sim \phi\left(\mathbf{x}_j' \boldsymbol{\beta}_i, \sigma_i\right), \quad \forall j \in \mathscr{C}_i,\ i = 1, \ldots, k, \tag{1}$$

where in many situations we can assume $\phi(\mathbf{x}_j'\boldsymbol{\beta}_i, \sigma_i)$ to be a normal density with mean $\mathbf{x}_j'\boldsymbol{\beta}_i$ and variance σ_i^2. This is equivalent to describing the data by a group of linear models:

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

Since the partition Π is unknown and the number of possible such partitions depends on n, the model (2) is a nonparametric one. Actually, it can be proved that the total number of nondegenerate partitions of form Π is equal to the Stirling number of the second kind $S(n,k) = (1/k!)\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$; confer Tomescu [18]. Also, the linear regression function in (2) can be extended to a nonlinear one including spline and local polynomial regression and so forth under this regression clustering setting. This extension will not be pursued in this paper. Further, the true distribution of e_j need not be the normal. Namely, we use $N(0, \sigma_i^2)$ only as a "working" distribution for e_j. Then the corresponding least squares or maximum likelihood approach becomes the quasi-likelihood one that still possesses many optimality properties (cf. chapter 9 of [19]). We resort to using a robust approach in this paper instead to deal with the violation of normality assumption.

Given the regression clustering model introduced above, we need to estimate the parameters $(\boldsymbol{\beta}_i, \sigma_i^2)_{i=1,\ldots,k}$ and find the best partition Π together with k for application. Optimal parameter estimation and partition can be achieved using the maximum likelihood principle, while finding the optimal k can be done based on an information criterion. The latter will be explained in next section. Now, we proceed to do parameter estimation and partition.

Under the fixed partition model (2), the log-likelihood function is given by

$$\log L_n\left(k, \Pi, \left(\boldsymbol{\beta}_1, \sigma_1^2\right), \ldots, \left(\boldsymbol{\beta}_k, \sigma_k^2\right)\right) = -\frac{1}{2}\sum_{i=1}^{k}\sum_{j\in\mathscr{C}_i}\left(\log 2\pi + \log\sigma_i^2 + \frac{\left(y_j - \boldsymbol{\beta}_i'\mathbf{x}_j\right)^2}{\sigma_i^2}\right). \tag{3}$$

It is clear that the best estimates of the parameters and the partition should be those maximizing the log-likelihood (3) for given k. However, the number of possible partitions $S(n,k)$ is astronomic even for moderate n and k; for example, $S(20,3) = 580,606,446$ and $S(50,3) \approx 1.2 \times 10^{23}$. Therefore, it is almost impossible to find the global optimal partition by enumeration. Here, we propose an iterative estimation method to find local optimal estimates of $(\boldsymbol{\beta}_i, \sigma_i^2)_{i=1,\ldots,k}$ and Π for a given k. This method extends the exchange method of Späth [16, 17].

When fixing $(\boldsymbol{\beta}_i, \sigma_i^2)_{i=1,\ldots,k}$ at given estimates $(\widehat{\boldsymbol{\beta}}_i, \widehat{\sigma}_i^2)_{i=1,\ldots,k}$, (3) achieves the maximum if each data point j belongs to cluster

$$\widehat{\mathscr{C}}_i = \arg\min_{1\le i\le k}\left(\log\widehat{\sigma}_i^2 + \frac{\left(y_j - \widehat{\boldsymbol{\beta}}_i'\mathbf{x}_j\right)^2}{\widehat{\sigma}_i^2}\right). \tag{4}$$

At given $\widehat{\mathscr{C}}_i$, $i = 1, \ldots, k$, (3) is the sum of the usual log-likelihood functions for homogeneous linear regressions within clusters. Hence, it is maximized at the least squares estimates $\widehat{\boldsymbol{\beta}}_i$ obtained based on the data points within $\widehat{\mathscr{C}}_i$, and

$$\widehat{\sigma}_i^2 = \frac{\sum_{j\in\widehat{\mathscr{C}}_i}\left(y_j - \widehat{\boldsymbol{\beta}}_i'\mathbf{x}_j\right)^2}{\widehat{n}_i}, \quad \text{where } \widehat{n}_i = \left|\widehat{\mathscr{C}}_i\right| \text{ is the size of } \widehat{\mathscr{C}}_i, \ i = 1, \ldots, k. \tag{5}$$

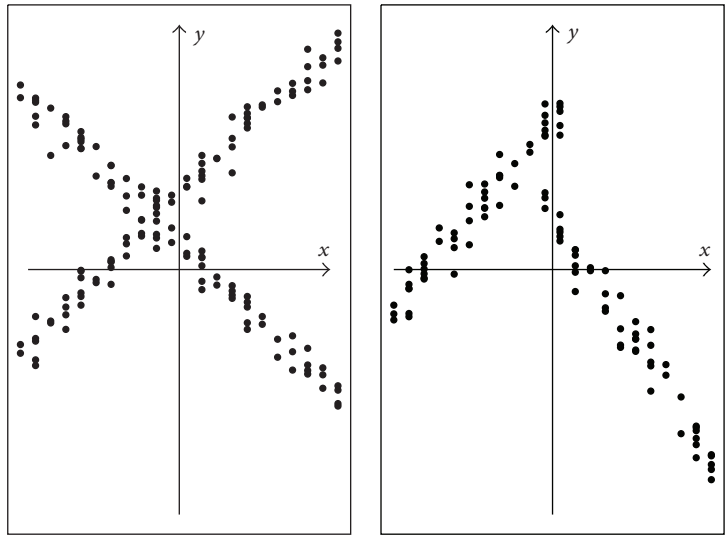
Then, $\log\widehat{L}_n$ is monotonically increased if the steps (4) and (5) are carried out alternately. This procedure leads to a local maximum in finitely many steps. It is expected to be a good approximation of the global maximum if an initial partition is properly chosen. In practice, we often assume that the variance parameters σ_i^2, $i = 1, \ldots, k$, have a common value σ^2 and estimate σ^2 by a pooled estimator. This modification tends to return a more robust partition than otherwise.

Note that the work in this section so far can be extended to multivariate regression clustering without any theoretical difficulty. The essential difference between multivariate regression and multiple regression is that the former has a vector response variable while the latter has a univariate one. Hence, (3) to (5) and the relevant ones in the rest of the paper can be easily modified to incorporate the vector response variable, from which it is ready to perform multivariate regression clustering. We will not get into the technical details involved but will provide a real data example in Section 6 to perform multivariate regression clustering.

It is well-known that the least squares method is very sensitive to outliers and violation of the normality assumption in the data. Robust methods can be developed to overcome this vulnerability. Among them, procedures based on M-estimation are considered here. M-estimation can be regarded as a generalization of the maximum likelihood estimation. A particular one is the maximum likelihood estimation based on Huber's least favourable distribution, whose density function is the normal at around the origin and the exponential in the tails. Using Huber's M-estimation method, we can drop the assumption $e_j \sim N(0, \sigma_i^2)$ in (2) and estimate $\boldsymbol{\beta}_i$ by minimizing $\sum_{j\in\widehat{\mathscr{C}}_i}\rho_c(y_j - \boldsymbol{\beta}_i'\mathbf{x}_j)$ for given partition $\widehat{\mathscr{C}}_i$, $i = 1, \ldots, k$. Here, $\rho_c(\cdot)$ is Huber's discrepancy function defined as

$$\rho_c(t) = \begin{cases} \frac{1}{2}t^2, & |t| < c, \\ c|t| - \frac{1}{2}c^2, & |t| \ge c, \end{cases} \tag{6}$$

where c is determined by the scale parameter in Huber's least favourable distribution. We find that assuming a constant scale parameter across all clusters tends to give better robust results, so we adopt this assumption in this paper. Now for given estimates $\widehat{\boldsymbol{\beta}}_i$, $i = 1, \ldots, k$, each data point j is assigned or reassigned to cluster $\widehat{\mathscr{C}}_i = \arg\min_{1\le i\le k}\rho_c(y_j - \widehat{\boldsymbol{\beta}}_i'\mathbf{x}_j)$. At this point, it can be seen that, instead of $\log\widehat{L}_n$, the function $\sum_{i=1}^{k}\sum_{j\in\mathscr{C}_i}\rho_c(y_j - \boldsymbol{\beta}_i'\mathbf{x}_j)$ will be monotonically increased if



(a) (b)

Figure 2: Assignment independence (a) versus assignment dependence (b).

the above two M-estimation steps are carried out alternately. This gives a robust counterpart of the likelihood-based local optimal estimation and selection introduced earlier in this section.

To conclude this section, note that the fixed partition approach has a particular advantage over the random partition one in the context of regression clustering or cluster regression. As observed by Hennig [20], the mixture probability model involved in random partitioning presumes implicitly an *assignment independence* of each object to clusters with respect to the covariate vectors $\mathbf{x}_j$. That is, the clusters keep the same proportions $\{\pi_i,\ i = 1, \ldots, k\}$ for every fixed covariate vector $\mathbf{x}_j$. In other words, the probability of a point $(y_j, \mathbf{x}_j')$ to be generated by cluster i is independent of $\mathbf{x}$ and j. This is generally not true as shown in Figure 2, which is adapted from Hennig [20]. On the other hand, the fixed partition model (2) supposes that the cluster membership of each object or cluster labels are explicitly parameterized and are determined by the estimation of $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\sigma}_i^2$ through the points $(y_j, \mathbf{x}_j')$, $j \in \mathscr{C}_i$. Hence, the fixed partition model does take care of the problem of possible *assignment dependence* between the jth object and the associated covariate $\mathbf{x}_j$. In principle, the random partition approach can be generalized to account for the assignment dependence, for example, by allowing $\{\pi_i,\ i = 1, \ldots, k\}$ to depend on $\mathbf{x}_j$. But the resultant probability model will be much more difficult to be analyzed both algebraically and numerically; and no such study can be found in literature so far to our knowledge.

3. Estimating the Number of Clusters

The number of clusters to be used in regression clustering is normally unknown so it should also be estimated. In this section we provide two procedures for estimating the number of clusters, one based on least squares estimation and the other on robust M-estimation.

We use a more detailed notation $\mathcal{O}^{(n)} = \{1, 2, \ldots, n\}$ to denote the n data objects which have observations $(y_1, \mathbf{x}_1')$, $\ldots, (y_n, \mathbf{x}_n')$ as described in previous sections. Recall that 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 subpopulations, each of which is characterized by a regression hyperplane with class-specific unknown parameters. Therefore, for the n observations from this population, there exists an underlying partition $\Pi_{k_0}^{(n)} = \{\mathcal{O}_1^{(n)}, \ldots, \mathcal{O}_{k_0}^{(n)}\}$, and by (2) each cluster $\mathcal{O}_i^{(n)} \triangleq \{i_1, \ldots, i_{n_i}\} \subseteq \mathcal{O}^{(n)}$ follows the regression model

$$\mathbf{y}_{\mathcal{O}_i} = X_{\mathcal{O}_i}\boldsymbol{\beta}_{0i} + \mathbf{e}_{\mathcal{O}_i}, \quad \mathbf{e}_{\mathcal{O}_i} \sim N\left(0, \sigma_i^2 I_{n_i}\right), \tag{7}$$

where $\mathbf{y}_{\mathcal{O}_i} = (y_{i_1}, \ldots, y_{i_{n_i}})'$, $X_{\mathcal{O}_i} = (\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{n_i}})'$ is an $n_i \times p$ design matrix in the cluster $\mathcal{O}_i$, $\mathbf{e}_{\mathcal{O}_i}$ is an n_i-vector of random errors, I_{n_i} is an $n_i \times n_i$ identity matrix, and $n_i = |\mathcal{O}_i|$ for $i = 1$, $\ldots, k_0$. Here, $(\boldsymbol{\beta}_{0i}', \sigma_i)' \in \mathbb{R}^p \times \mathbb{R}^+$, $1 \le i \le k_0$, are k_0 unknown parameter vectors; and $\boldsymbol{\beta}_{0i}$, $1 \le i \le k_0$, are assumed to be distinct from one another. It is clear that $n = n_1 + \cdots + n_{k_0}$. In the following, we assume that $k_0 \le K$, where K is a known positive integer. Note that in (7) we have suppressed the n in $\mathcal{O}_i^{(n)}$ for simplicity of presentation. Also note that the normality assumption for the random errors $\mathbf{e}_{\mathcal{O}_i}$, although reasonable in many situations, is just a "working" distribution and not really required for applying the least squares method.

In order to estimate k_0, we fit a regression clustering model to the data for each $k \le K$ using the methods developed in Section 2. A criterion function of k can be obtained from the cluster regression fitting. Then k_0 is estimated as the minimizer of the criterion function. Shao

and Wu [21] have used this idea to develop an information-based criterion for estimating k_0. Let $\Pi_k^{(n)} = \{\mathscr{C}_1^{(n)}, \ldots, \mathscr{C}_k^{(n)}\}$ be an arbitrary k-cluster partition of the n observations. Shao and Wu's information criterion is defined as

$$D_n\left(\Pi_k^{(n)}\right) = \sum_{i=1}^{k} \left\| \mathbf{y}_{\mathscr{C}_i^{(n)}} - X_{\mathscr{C}_i^{(n)}} \widehat{\boldsymbol{\beta}}_i \right\|^2 + q(k) A_n, \tag{8}$$

where $q(k)$ is a strictly increasing positive function of k, A_n is a sequence of positive constants, $\widehat{\boldsymbol{\beta}}_i$ are least squares estimators, and $\|\cdot\|$ is the Euclidean norm. Typically, $q(k) = kp$ and $A_n \propto \log(n)$ or $A_n \propto \log\log(n)$ are chosen. Then $\widehat{k}_n$, the estimate of k_0, is the integer that minimizes this criterion, that is,

$$D_n\left(\widehat{k}_n\right) = \min_{1 \le k \le K} \min_{\Pi_k^{(n)}} D_n\left(\Pi_k^{(n)}\right). \tag{9}$$

It can be seen that in (8) the first term is the sum of residual squares which measures the goodness of fit of the model and the second term is the penalty for overfitting. Moreover, the criterion (9) shows that one determines the optimal number of clusters and the corresponding partitioning simultaneously. We shall call (8) together with (9) 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 [21] that the proposed Criterion LS-C 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.

Concerning the robustness of regression clustering, one can use a robust criterion to estimate the underlying number of clusters k_0, where we assume that each cluster $\mathscr{O}_i \triangleq \{i_1, \ldots, i_{n_i}\} \subseteq \mathscr{O}^{(n)}$ is characterized by a linear model:

$$y_{j,\mathscr{O}_i} = \mathbf{x}'_{j,\mathscr{O}_i} \boldsymbol{\beta}_{0i} + e_{j,\mathscr{O}_i}, \quad j \in \mathscr{O}_i, \tag{10}$$

with the random error $e_{j,\mathscr{O}_i}$ not following any specific distribution contrary to that in the linear model (7). In particular, Rao et al. [22] have developed the following robust information criterion function for estimating k_0:

$$R_n\left(\Pi_k^{(n)}\right) = \sum_{s=1}^{k} \sum_{j \in \mathscr{C}_s} \rho_c\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s} \widehat{\boldsymbol{\beta}}_s\right) + q(k) A_n, \tag{11}$$

where ρ_c is Huber's discrepancy function and $\widehat{\boldsymbol{\beta}}_s$ are the M-estimators described in Section 2 or equivalently satisfying

$$\begin{aligned} &\sum_{j \in \mathscr{C}_s} \rho_c\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s} \widehat{\boldsymbol{\beta}}_s\right) \\ &\quad = \min_{\boldsymbol{\beta}_s} \sum_{j \in \mathscr{C}_s} \rho_c\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s} \boldsymbol{\beta}_s\right). \end{aligned} \tag{12}$$

It can be seen that similar to that in (8) the first term in (11) is a generalization of a minimum negative log-likelihood function derived from Huber's least favourable distribution, and the second term is the penalty for overfitting.

Using (11), the estimate $\widehat{k}_n$ of the underlying number of clusters k_0 is the one satisfying

$$R_n\left(\widehat{k}_n\right) = \min_{1 \le k \le K} \min_{\Pi_k^{(n)}} R_n\left(\Pi_k^{(n)}\right). \tag{13}$$

We shall call (11) together with (13) Criterion RM-C, which stands for the clustering based on robust M-estimation. Similar to Criterion LS-C, Criterion RM-C implies that one determines the optimal number of clusters and the corresponding partitioning simultaneously.

In Rao et al. [22], it is shown that the true clustering and the associated regression hyperplanes are attained with probability 1 by RM-C when n increases to infinity and under certain mild conditions. In particular, normal distribution assumption is not required for the random errors in each regression cluster.

4. Pointwise Iterative Algorithms for Regression Clustering Estimation, Partition, and Selection

Computing algorithms can be written to implement the regression clustering methods described in Sections 2 and 3. Recall that in the methods we first estimate the optimal partition $\Pi_k = \{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$ and the regression parameters simultaneously by minimizing certain within-cluster sum of residual squares sums (SRSS) or alike for each fixed k. The quantity to be minimized is equivalent to

$$\mathrm{SRSS}\left(\Pi_k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\right) = \sum_{i=1}^{k} \left\| \mathbf{y}_{\mathscr{C}_i} - X_{\mathscr{C}_i} \boldsymbol{\beta}_i \right\|^2 \tag{14}$$

for LS regression clustering or sum of robust residual squares sums (RRSS)

$$\mathrm{RRSS}\left(\Pi_k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\right) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \rho_c\left(y_{j,\mathscr{C}_i} - \mathbf{x}'_{j,\mathscr{C}_i} \boldsymbol{\beta}_i\right) \tag{15}$$

for an M-estimation based robust regression clustering. Only local minimization results can be guaranteed here. We process this local minimization for each candidate k and use Criterion LS-C or RM-C to determine the best k. The whole procedure can be accomplished according to the following algorithm:

(i) Label all the observations from 1 to n (order does not matter). Given an initial partition $\Pi_k = \{\mathscr{C}_1, \ldots, \mathscr{C}_k\}$ of $\mathscr{O} = \{1, \ldots, n\}$, fit a regression model (or a robust regression model with a $\rho_c(\cdot)$ function for RM-C criterion) in each of the k clusters and obtain the sum of the residual squares sums SRSS_0 or RRSS_0 for this partition. Let $i = 0$.

(ii) Set $i = i + 1$, and reset $i = 1$ if $i > n$. Identify $\mathscr{C}_j$ such that $i \in \mathscr{C}_j$. Then move i into $\mathscr{C}_h$ with $h = 1, \ldots, k$ and $h \neq j$, respectively. For each of these $k - 1$ relocations, refit the model by regression clustering (or robust regression clustering) and calculate the sum of the

residual squares sums (or RRSS) accordingly. Denote the smallest one by $SRSS_h$ or $RRSS_h$. If $SRSS_h < SRSS_0$ (or $RRSS_h < RRSS_0$ in robust procedure), redefine $\mathscr{C}_j = \mathscr{C}_j - \{i\}$ and $\mathscr{C}_h = \mathscr{C}_h + \{i\}$, and set $SRSS_0 = SRSS_h$ (or $RRSS_0 = RRSS_h$). Otherwise, return to the beginning of (ii).

(iii) Repeat (ii) until the objective function (14) or (15) does not decrease any further, which means that no observation relocation is necessary and the optimal clustering is achieved for this k.

(iv) Proceed with (i) to (iii) for each candidate k and use the Criterion LS-C or RM-C to find $\widehat{k}_n$, the optimal number of clusters.

It is important to use a good initial partition of $\{1, \ldots, n\}$ in running steps (i) to (iii) so that the global minimum of (14) or (15), or its good approximation, can be achieved. We propose to generate the initial partition of a dataset using the following algorithm which we find works well in practice:

(I1) Consider the linear model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i. \tag{16}$$

Based on the whole dataset, one estimates $\boldsymbol{\beta}$ by a robust method, for example, least median regression or least trimmed squares method [23]. Note that a random seed is implicitly used in such robust methods.

(I2) Put into set C_1 those data points whose distances to the regression hyperplane estimated in Step (I1) are less than a predetermined number, say δ. 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 (I5). Here, C_1^c is the complementary set of C_1.

(I3) Based on the dataset $\bigcap_{i=1}^{\ell} C_i^c$, one estimates $\boldsymbol{\beta}$ in (16) by the same robust method used in Step (I1).

(I4) Put into $C_{\ell+1}$ those points in $\bigcap_{i=1}^{\ell} C_i^c$ whose distances to the regression hyperplane estimated in Step (I3) are less than δ. 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 (I3); otherwise, go to Step (I5).

(I5) 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 $\ell = 0$.

One can adjust the values of δ and m either in advance or adaptively to get an initial partition of k clusters for any given k. For example, set m to the integer part of $0.5n/k$ and δ to the best value such that two clusters can be obtained in (I2).

The above initial partition algorithm gives essentially an iterated hierarchical binary clustering method, where each binary clustering is realized through resistant regression such as the least median regression. The resistant regression is robust, having high breakdown threshold; thus, although not being fully efficient, it is highly likely to produce a reasonable initial partition through its iterated executions.

Table 2: Confusion matrix between `sepal.length` ~ `sepal.width` regression clusters and species information.

	setosa	*versicolor*	*virginica*
Cluster 1	50	1	1
Cluster 2	0	35	16
Cluster 3	0	14	33

Table 3: Confusion matrix between `sepal.length` ~ `sepal.width` + `petal.length` + `petal.width` regression clusters and species information.

	setosa	*versicolor*	*virginica*
Cluster 1	25	17	14
Cluster 2	12	19	15
Cluster 3	13	14	21

The two algorithms consisting of Steps (I1) to (I5) and (i) to (iv) may be named IPARC to reflect the iterative pointwise assessing nature in regression clustering.

5. Example and Simulation Study

In this section, we first apply regression clustering to the *Iris* data and provide a brief guideline on when to use the method properly. We then present a simulation study to assess the finite sample performance of Criteria LS-C and RM-C.

5.1. The Iris Data Example. Recall the *Iris* data that we analyzed using the k-means method in Section 1. Now we want to use the regression relationship between sepal length and sepal width variables to partition the 150 observations of sepal length and width and petal length and width into 3 clusters. Statistics package R is used to implement our IPARC procedure, where we set $m = 2p$ and $\delta = 0.05$ or 0.2 and initial random seed being determined by `set.seed(123456)`, and use only the least squares estimation in this example. The partition result and its comparison with the species information are summarized in Table 2. Comparing Tables 1 and 2, we see that the cluster information revealed by the cluster regression `sepal.length` ~ `sepal.width` is very much the same as that by the k-means and conforms with the species information.

When we use cluster regression `sepal.length` ~ `sepal.width` + `petal.length` + `petal.width` to partition the data into 3 clusters, we get a result summarized in Table 3 which is very different from Tables 1 and 2. This confirms that the cluster label information obtained from regression clustering has a different interpretation from that obtained from the k-means. The former tells us how differently regression performs across the clusters, while the latter tells us how distances among data observations themselves behave differently across the clusters. Fitting this regression clustering to the data gives SRSS = 2.901 and coefficients of determination of 0.972, 0.958, and 0.971, respectively, for the 3 regression hyperplanes. On the other hand, when we fit the same regression model to the 3 clusters determined by the k-means, we get SRSS = 12.699

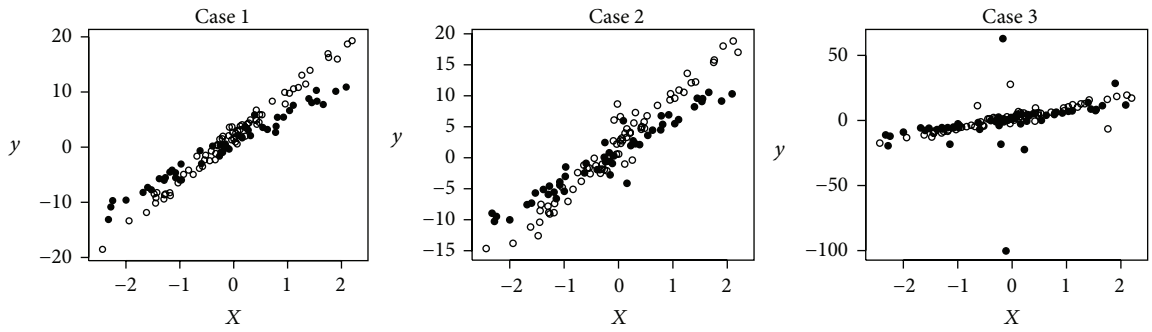


Figure 3: Simulated data with two clusters.

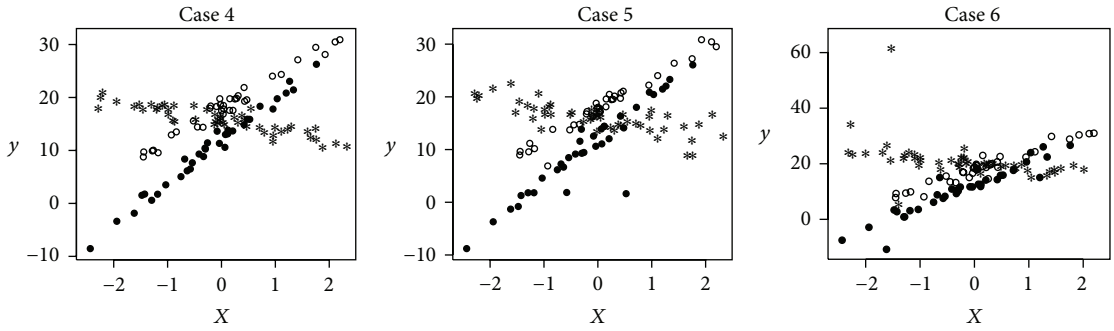


Figure 4: Simulated data with three clusters.

Table 4: Shorthand notation for six cases.

N1C2	Case 1	Two regression lines	Normal error
T1C2	Case 2	Two regression lines	$t(3)$ error
C1C2	Case 3	Two regression lines	Cauchy(0, 1) error
N1C3	Case 4	Three regression lines	Normal error
T1C3	Case 5	Three regression lines	$t(3)$ error
C1C3	Case 6	Three regression lines	Cauchy(0, 1) error

and coefficients of determination of 0.575, 0.525, and 0.578. Similar results are obtained if the same regression model is fit to the 3 clusters determined by the species variable. Therefore, regression clustering method is fundamentally different from the general cluster analysis methods such as the k-means. One should use regression clustering if partitioning data to conform to the regression pattern is of interest.

5.2. Simulation Study. We use simulated data sets to assess the finite sample performance of Criteria LS-C and RM-C for regression clustering. Two factors will be considered for this simulation: number of clusters (2 or 3) and error distributions ($N(0, 1)$, $t(3)$, or Cauchy(0, 1)), so there are in total 6 cases of data to be considered, which are summarized in Table 4. There will be only one covariate involved in the regression clustering and the covariate is generated from $N(0, 1)$. The parameters used for each case are given in Table 5. Then, the fixed partition regression clustering model $y_{ji} = \mathbf{x}'_{ji}\boldsymbol{\beta}_{0i} + e_{ji}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k_0$, is applied to generate the response values y_{ji}, where e_{ji} is a random number originating from $N(0, 1)$, $t(3)$, or Cauchy(0, 1), and the first element of $\mathbf{x}_{ji}$ is the constant 1 corresponding to the intercept term in the model.

Figures 3 and 4 give us an intuition of what the data typically look like for Cases 1–6 with normal, $t(3)$, or Cauchy 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. The groupings are hard to see with Cauchy(0, 1) random errors.

In this study, we set $q(k) = kp$, where p is the number of regression coefficients in the model and is a constant in our study; k is the unknown number of clusters that we are seeking. It is noted that in an information model selection criterion, a penalty function, which is A_n in (8) or (11), is usually chosen as $c \log(n)$ or $c \log\log(n)$ with a constant $c > 0$. In light of the fact that $\lim_{\lambda \to 0}[(\log n)^{\lambda} - 1]/\lambda = \log\log n$, we set $A_n = [(\log n)^3 - 1]/3$.

The ρ_c function we employed for M-estimation is $\rho_c(u) = 0.5u^2$ if $|u| \leq 1.345$ and $\rho_c(u) = 1.345|u| - 0.5 \times 1.345^2$ otherwise (Huber ρ_c). In the following, when we state the simulation results, Criterion RM-C means M-estimation based regression clustering procedure with Huber's ρ_c exclusively.

For each of the six cases, we conduct 1000 simulations using Criteria LS-C and RM-C separately. To apply the algorithm IPARC, we set $\delta = 0.2$ and $m = 2p$. The algorithm given in the previous section is then used to estimate the number of clusters in cluster linear regression. In Tables 6 and 7, we summarize the results from the simulation study, where each number represents the relative frequencies of selecting the possible numbers of clusters k out of 1000 replications.

Table 5: Parameter values used in the simulation study of regression clustering.

Case	k_0	Regression coefficients	Number of observations
1–3	2	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 2 \\ 8 \end{pmatrix}, \boldsymbol{\beta}_{02} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}$	$n_1 = 70, n_2 = 50$
4–6	3	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 18 \\ 6 \end{pmatrix}, \boldsymbol{\beta}_{02} = \begin{pmatrix} 12 \\ 8 \end{pmatrix}, \boldsymbol{\beta}_{03} = \begin{pmatrix} 15 \\ -2 \end{pmatrix}$	$n_1 = 35, n_2 = 35, n_3 = 50$

Table 6: Relative frequencies of selecting k based on 1000 simulations for Cases 1–3.

$k_0 = 2$	Case 1 ($N(0, 1)$ error)		Case 2 ($t(3)$ error)		Case 3 (Cauchy(0, 1) error)	
	LS-C	RM-C	LS-C	RM-C	LS-C	RM-C
$k = 1$	0.000	0.000	0.001	0.001	0.005	0.006
$k = 2$	0.986	1.00	0.422	0.999	0.292	0.745
$k = 3$	0.014	0.000	0.488	0.000	0.415	0.183
$k = 4$	0.000	0.000	0.087	0.000	0.227	0.055
$k = 5$	0.000	0.000	0.002	0.000	0.061	0.011

Table 7: Relative frequencies of selecting k based on 1000 simulations for Cases 4–6.

$k_0 = 3$	Case 4 ($N(0, 1)$ error)		Case 5 ($t(3)$ error)		Case 6 (Cauchy(0, 1) error)	
	LS-C	RM-C	LS-C	RM-C	LS-C	RM-C
$k = 1$	0.000	0.000	0.000	0.000	0.000	0.000
$k = 2$	0.000	0.000	0.000	0.002	0.117	0.012
$k = 3$	1.00	1.00	0.791	0.997	0.232	0.611
$k = 4$	0.000	0.000	0.207	0.001	0.566	0.350
$k = 5$	0.000	0.000	0.002	0.000	0.085	0.027

It is clear that Criterion LS-C performs almost perfectly in Cases 1 and 4 since the errors are standard normal distributed. However, when there exist outliers in the data set or the normality of the data is violated, Criterion LS-C performs poorly. On the contrary, as shown in Tables 6 and 7, Criterion RM-C does as nearly perfect a job as Criterion LS-C in Cases 1 and 4; at the same time, neither outliers nor abnormality has much effect on its ability to detect the underlying true number of regression hyperplanes in the data.

In addition to the robustness shown above in selecting the number of clusters, the procedure of the M-estimation based regression clustering is also robust in partitioning the data. Table 8 presents the estimation of the regression parameters by applying LS-C and RM-C to the data shown in Figures 3 and 4. From the table, it can be seen that when the errors are $t(3)$ or Cauchy(0, 1) distributed, the LS regression clustering method is not able to capture the underlying groupings, while the M-estimation based regression clustering method detects the true linear patterns in the data, in spite of the abnormality in the data.

6. Analysis of RGB Cell Tracking Data

Recently, a new technique called RGB marking has been introduced to facilitate the identification of individual cell clones in both in vivo and in vitro experiments [24]. RGB marking introduces three lentiviral vectors in individual cells encoding the basic colors red, green, and blue. Raw image data representing 128 colorectal cancer cells are shown in Figure 5; the same data are to be analyzed in detail in this section. Since the colored cells are easily identifiable within whole organ structures, scientists can track the cells and determine their role during processes such as organ regeneration, malignant outgrowth, or immune responses.

To this end, scientists are required to cluster cell types according to some basic color combinations. Due to the variability of the vector insertion, however, single RGB-marked cells express fluorescent proteins at different and very characteristic levels. The underlying principle of additive color mixing, similar to that in computer or TV screens, generates different color combinations that can be used to discriminate individual cell clones. The main difficulty in this kind of data is that the intrinsic variability of the underlying biological mechanisms makes the actual number of distinguishable colors generated by RGB marking in a tissue difficult to predict. In addition, cell intensities for different colors are known to vary depending on the cell area, which is an indicator of cell morphology.

The data set analyzed in this section consists of measurements on colorectal cancer cell lines expressing various quantities of three different fluorescent proteins: Cerulean (blue), Venus (yellow/green), and mCherry (red). The genes

Table 8: The estimation of the regression parameters by applying LS-C and RM-C to the data shown in Figures 3 and 4.

k_0	Case	Clusters	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$
2	1	True	$\begin{pmatrix}2\\8\end{pmatrix}$	$\begin{pmatrix}1\\5\end{pmatrix}$		
		LS-C	$\begin{pmatrix}2.12\\8.02\end{pmatrix}$	$\begin{pmatrix}0.76\\5.11\end{pmatrix}$		
		RM-C	$\begin{pmatrix}2.11\\8.03\end{pmatrix}$	$\begin{pmatrix}0.78\\5.10\end{pmatrix}$		
	2	LS-C	$\begin{pmatrix}1.48\\5.56\end{pmatrix}$	$\begin{pmatrix}-1.13\\5.87\end{pmatrix}$	$\begin{pmatrix}4.46\\6.18\end{pmatrix}$	
		RM-C	$\begin{pmatrix}2.21\\7.89\end{pmatrix}$	$\begin{pmatrix}0.73\\4.89\end{pmatrix}$		
	3	LS-C	$\begin{pmatrix}2.40\\6.66\end{pmatrix}$	$\begin{pmatrix}-46.33\\-11.23\end{pmatrix}$		
		RM-C	$\begin{pmatrix}2.29\\8.42\end{pmatrix}$	$\begin{pmatrix}0.59\\5.17\end{pmatrix}$		
3	4	True	$\begin{pmatrix}18\\6\end{pmatrix}$	$\begin{pmatrix}12\\8\end{pmatrix}$	$\begin{pmatrix}15\\-2\end{pmatrix}$	
		LS-C	$\begin{pmatrix}18.05\\6.06\end{pmatrix}$	$\begin{pmatrix}11.97\\8.02\end{pmatrix}$	$\begin{pmatrix}14.66\\-1.85\end{pmatrix}$	
		RM-C	$\begin{pmatrix}18.04\\6.07\end{pmatrix}$	$\begin{pmatrix}11.95\\8.03\end{pmatrix}$	$\begin{pmatrix}14.66\\-1.87\end{pmatrix}$	
	5	LS-C	$\begin{pmatrix}17.74\\6.14\end{pmatrix}$	$\begin{pmatrix}12.02\\8.16\end{pmatrix}$	$\begin{pmatrix}10.73\\-2.87\end{pmatrix}$	$\begin{pmatrix}15.54\\-1.70\end{pmatrix}$
		RM-C	$\begin{pmatrix}17.88\\5.98\end{pmatrix}$	$\begin{pmatrix}12.14\\8.14\end{pmatrix}$	$\begin{pmatrix}14.94\\-1.88\end{pmatrix}$	
	6	LS-C	$\begin{pmatrix}18.23\\6.29\end{pmatrix}$	$\begin{pmatrix}12.28\\8.27\end{pmatrix}$	$\begin{pmatrix}15.20\\-2.10\end{pmatrix}$	$\begin{pmatrix}32.17\\-27.23\end{pmatrix}$
		RM-C	$\begin{pmatrix}18.02\\6.26\end{pmatrix}$	$\begin{pmatrix}12.24\\7.99\end{pmatrix}$	$\begin{pmatrix}15.19\\-2.09\end{pmatrix}$	

coding for the fluorescent proteins were transferred into the cells via lentivirus-mediated transduction at a less than 100% efficiency so that most cells expressed different quantitative combinations of all three fluorescent proteins as described by Weber et al. [24]. The cells were imaged on a high-content imager (Operetta, Perkin Elmer). The final data consisted of fluorescent intensities of red, blue, and green color channels (electromagnetic wavelength in nanometers, nm), morphology parameters including cell areas, and spatial coordinates for 128 cells.

Figure 5 shows the original data and clustering obtained by the LS regression clustering approach defined in (14), using multivariate regression with color intensities as the response vector, with morphological predictor (cell area) being used in (c), and without using any predictors in (d). Clustering methods are relatively robust to the initial random seed (here we used `set.seed(111)`) in both cases. When the cell area predictor is included, the resulting clustering changes, thus suggesting that the cell morphology information (cell area) plays a role in separating different cells types. In Table 9, we summarize the outcome for this LS regression clustering.

To select the optimal number of clusters, we used the information criterion function (8) for LS and (11) for RM, with $q(k) = k$, where k is the unknown number of clusters that we are seeking for. Figure 6 shows the optimal numbers of clusters using $A_n = \log\log n$ (C1) and $A_n = \log n$ (C2) for both clustering approaches. Robust clustering is carried out using Huber's discrepancy function (6) with the tuning constant $c = 1.345$ being chosen. The resulting optimal number of clusters based on C1 is 5 by both LS and RM regression clustering criteria, which is compatible with biological considerations.

Finally, we assess the performances of the LS and RM regression clustering and compare them with that of the

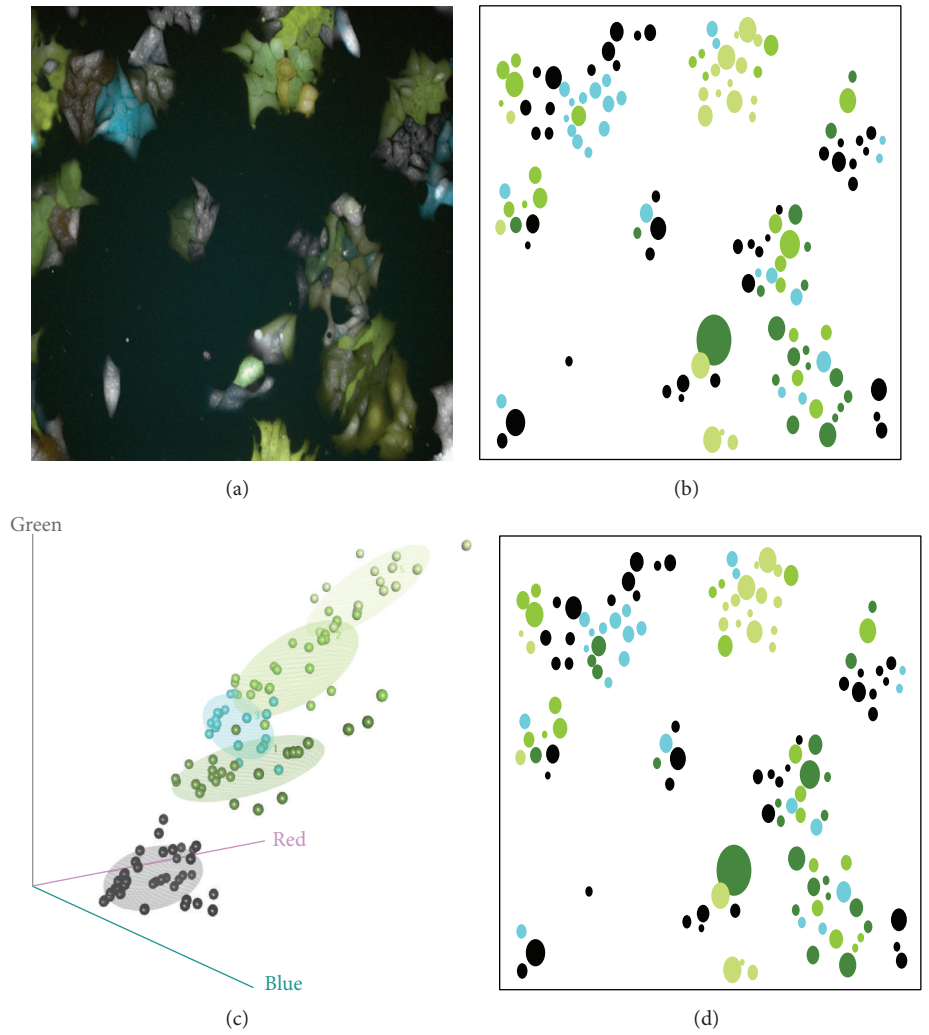


Figure 5: (a) Raw spatial data on 128 colorectal cancer cells imaged on a high-content microscope imager (Operetta, Perkin Elmer). (b) Spatial distribution of the 128 cells represented by the colored circles. The circles have 5 colors representing 5 clusters which resulted from LS multivariate regression clustering minimizing (14). The size of each circle estimates the area of the corresponding cell. The clustering uses RGB intensities as the response vector and cell area as the predictor. (c) 3D-scatterplot of the clustered RGB intensities of the 128 cells. Colors of the points show the same 5 clusters shown in (b). (d) Spatial distribution of the 128 cells, with the colored circles showing the 5 clusters given by the LS multivariate regression clustering not including any predictor.

Table 9: Summary statistics based on the 5 clusters obtained from the multivariate LS regression clustering including the cell area covariate: sample means and standard deviations of the 3-dimensional response vector (i.e., RGB intensities on log scale), as well as the number of observations (i.e., cells) in each cluster.

Cluster	Mean (red, green, and blue)	$\widehat{\text{SD}}$ (red, green, and blue)	Cluster size
1	(4.99, 5.02, 5.78)	(0.10, 0.17, 0.25)	25
2	(5.66, 5.83, 5.78)	(0.23, 0.26, 0.18)	23
3	(5.40, 5.36, 5.57)	(0.12, 0.18, 0.14)	20
4	(4.55, 4.19, 5.52)	(0.28, 0.16, 0.11)	41
5	(6.32, 6.50, 5.92)	(0.22, 0.24, 0.18)	19

k-means method. The prediction strength (PS) statistic introduced by Tibshirani and Walther [25] is used for the assessment.

For a candidate number of clusters k ($k = 5$ in our case), let $\widehat{\mathscr{C}}_{\text{te}} = \{\widehat{\mathscr{C}}_{\text{te},1}, \ldots, \widehat{\mathscr{C}}_{\text{te},k}\}$ denote the partition of the test set resulting from regression clustering on all the data. Let $n_1, \ldots, n_k$ be the number of observations in these clusters. Let $\widehat{\mathscr{C}}_{\text{tr}}$ be the partition of the test set resulting from regression clustering on the training set. In particular, in the latter case each data point in the test set is clustered using (4) with $\widehat{\beta}_i$, $i = 1, \ldots, k$, produced by the training set.

Following notations of Tibshirani and Walther [25], denote $D[\widehat{\mathscr{C}}_{\text{tr}}, \widehat{\mathscr{C}}_{\text{te}}]$ as the $n \times n$ comembership matrix, with ii'th element $D[\widehat{\mathscr{C}}_{\text{tr}}, \widehat{\mathscr{C}}_{\text{te}}]_{ii'} = 1$, if a pair of observations i and

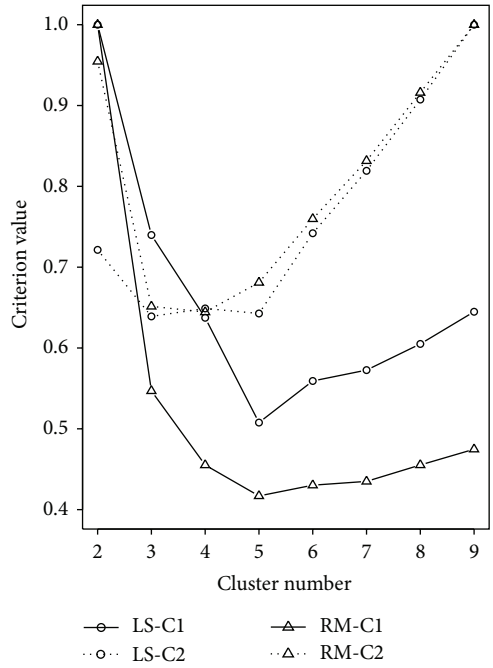


Figure 6: Model selection based on information criterion (8) with A_n equal to $\log\log n$ (C1) and $\log n$ (C2) for least squares (LS) and robust M-estimation (RM) based regression clustering approaches, respectively. All criterion values are scaled between 0 and 1.

i' that belong to the same cluster in $\widehat{\mathscr{C}}_{\text{te}}$ (i.e., $i \neq i' \in \widehat{\mathscr{C}}_{\text{te},j}$, $j = 1, \ldots, k$) also fall into the same cluster in $\widehat{\mathscr{C}}_{\text{tr}}$, and 0 otherwise. The prediction strength statistic can be written as

$$\text{PS} = \min_{1\leq j\leq k} \frac{1}{n_j\left(n_j - 1\right)} \sum_{i\neq i'\in\widehat{\mathscr{C}}_{\text{te},j}} D\left[\widehat{\mathscr{C}}_{\text{tr}}, \widehat{\mathscr{C}}_{\text{te}}\right]_{ii'}. \quad (17)$$

Therefore, the prediction strength is the proportion of observation pairs in the worst performing test cluster whose clustering results remain unchanged when clustering them by the training set clustering rule. Clearly, a regression clustering result has higher predictive power if the associated PS is higher.

For our data, we assess the clustering performance by cross-validation using 4 random partitions of our sample. Cross-validated prediction strength values for k-means, LS, and RM regression clustering methods are 0.44, 0.80, and 0.66, respectively. This suggests that the LS regression clustering is superior to the k-means. Moreover, due to the absence of strong deviations from the multivariate normal model for these data, the out-of-sample prediction strength of the LS regression clustering is larger than that of the robust RM regression clustering approach.

7. Conclusion

In this paper, we review the general cluster analysis methods and then focus on regression clustering which uses the model based fixed partition method and clusters the data based on the dependence between the response and explanatory variables. We provide both least squares based and robust M-estimation based methods for estimating parameters, partitioning the data, and selecting the optimal number of clusters in regression clustering. Algorithms have been developed to implement these methods. The example and simulation study conclude a satisfactory finite sample performance of the algorithms. Applying our developed method to regression cluster the RGB cells tracking data gives results compatible with biological considerations. It is known that the methods can only provide a local optimization solution and are computing intensive especially when the sample size is large. Currently, we are investigating these issues and expect to provide an improved solution to be reported elsewhere in the near future.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank Christina Mølck from the Department of Pathology at the University of Melbourne for carrying out the RGB cell marking experiment and Cameron Nowell from the Institute of Pharmaceutical Sciences at Monash University for doing the high-content imaging and producing the raw data that is analyzed in Section 6. Further, the authors would like to thank Louis Vermeulen and Maartje van der Heijden (University of Amsterdam, the Netherlands) for their generous gift of DLD1-LeGO cells as used in Section 6.

References

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

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam and J. Neyman, Eds., vol. 1, pp. 281–297, University of California Press, 1967.

[3] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: a k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, Heidelberg, Germany, 2nd edition, 2009.

[5] B. Clarke, E. Fokoué, and H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, Springer, Heidelberg, Germany, 2009.

[6] E. Anderson, "The irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, vol. 59, pp. 2–5, 1935.

[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[8] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015, http://www.R-project.org/.

[9] D. Pollard, "Strong consistency of k -means clustering," *The Annals of Statistics*, vol. 9, no. 1, pp. 135–140, 1981.

[10] J. A. Hartigan, "Consistency of single linkage for high-density clusters," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 388–394, 1981.

[11] M. A. Wong, "A hybrid clustering method for identifying high-density clusters," *Journal of the American Statistical Association*, vol. 77, no. 380, pp. 841–847, 1982.

[12] 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.

[13] 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.

[14] H. H. Bock, "The equivalence of two extremal problems and its application to the iterative classification of multivariate data," in *Manuscript for the Conference "Medizinische Statistik"*, Forschungsinstitut Oberworfach, 1969.

[15] 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 Publishing, River Edge, NJ, USA, 1996.

[16] H. Späth, "Algorithm 39 Clusterwise linear regression," *Computing*, vol. 22, no. 4, pp. 367–373, 1979.

[17] H. Späth, "Algorithm 48: a fast algorithm for clusterwise linear regression," *Computing*, vol. 29, no. 2, pp. 175–181, 1982.

[18] I. Tomescu, *Problems in Combinatorics and Graph Theory*, Wiley-Interscience, New York, NY, USA, 1985.

[19] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, London, UK, 2nd edition, 1989.

[20] C. Hennig, "Identifiability of models for clusterwise linear regression," *Journal of Classification*, vol. 17, no. 2, pp. 273–296, 2000.

[21] 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.

[22] C. R. Rao, 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*, vol. 2007, Article ID 37475, 15 pages, 2007.

[23] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, NY, USA, 1987.

[24] K. Weber, M. Thomaschewski, D. Benten, and B. Fehse, "RGB marking with lentiviral vectors for multicolor clonal cell tracking," *Nature Protocols*, vol. 7, no. 5, pp. 839–849, 2012.

[25] R. Tibshirani and G. Walther, "Cluster validation by prediction strength," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 511–528, 2005.