A New Method for Mining Regression Classes in Large Data Sets

Yee Leung, Jiang-Hong Ma, and Wen-Xiu Zhang

Abstract—Extracting patterns and models of interest from large databases is attracting much attention in a variety of disciplines. Knowledge discovery in databases (KDD) and data mining (DM) are areas of common interest to researchers in machine learning, pattern recognition, statistics, artificial intelligence, and high performance computing. An effective and robust method, coined regression-class mixture decomposition (RCMD) method, is proposed in this paper for the mining of regression classes in large data sets, especially those contaminated by noise. A new concept, called "*regression class*" which is defined as a subset of the data set that is subject to a regression model, is proposed as a basic building block on which the mining process is based. A large data set is treated as a mixture population in which there are many such regression classes and others not accounted for by the regression models. Iterative and genetic-based algorithms for the optimization of the objective function in the RCMD method are also constructed. It is demonstrated that the RCMD method can resist a very large proportion of noisy data, identify each regression class, assign an inlier set of data points supporting each identified regression class, and determine the a priori unknown number of statistically valid models in the data set. Although the models are extracted sequentially, the final result is almost independent of the extraction order due to a novel dynamic classification strategy employed in the handling of overlapping regression classes. The effectiveness and robustness of the RCMD method are substantiated by a set of simulation experiments and a real-life application showing the way it can be used to fit mixed data to linear regression classes and nonlinear structures in various situations.

Index Terms—Data mining, genetic algorithm, maximum likelihood method, mixture modeling, RCMD method, regression class, robustness.

1 Introduction

It is well-known that statistics is the art and science of extracting useful information and patterns of interest from empirical data. From this point of view, statistics, in a certain sense, is similar to data mining (DM) and knowledge discovery in databases (KDD) (see [12], [15]). Nonetheless, problems and methods of DM have some distinct features of their own. The most remarkable feature is that DM is concerned with structure discovery in large data sets. Although many problems in the field of DM have been tackled by techniques such as statistics, machine learning, pattern recognition, and artificial intelligence, developing effective mining methods for large data sets, especially contaminated by noise, is still an important and challenging problem [9], [11], [14], [16], [35].

Usually, the methods of extracting information from data sets are divided into two kinds in statistics: one is the general class of method using unlabeled samples, known as clustering (or "unsupervised learning"), which attempts to partition a set of observations by grouping them into a number of statistical classes; the other is classification (or "supervised learning") dealing with labeled samples. Both of them have been widely applied to a variety of disciplines such as computer vision, pattern recognition, remote sensing, marketing, and finance [2], [5].

One of the effective ways for conveying information is to use parametric stochastic models that can describe more completely information contained in a data set. Therefore, as a practice, we should first try to use parametric models in order to get more knowledge or information about a data set. Among parametric models, the parametric regression model can generally provide a more exact description of the underlying data and their quantitative interpretation. Furthermore, classification can be considered as a regression problem with the dependent variable taking discrete values. Data mining is often interested in simple and interpretative models. Thus, parametric regression is undoubtedly an appealing model for data analysis.

Naturally, we hope that a single regression model can be applied to a large or complicated data set if there exists such a model in it. Unfortunately, regression analysis is usually not appropriate for the study of large data sets, especially with noise contamination. The main reasons are as follows:

1. Regression analysis handles a data set as a whole. Even with the computer hardware available today, there are no effective means—such as processors and storage—for manipulating and analyzing a large amount of data.
2. More importantly, it is not real to assume that a single model can fit a large data set. It is highly likely that we need multiple models to fit a large data set. That is, a data set may not be accurately modeled using any single structure.
3. Classical regression analysis is based on stringent model assumptions. However, the real world, in

- *Y. Leung is with the Department of Geography, Center for Environmental Policy and Resource Management and Joint Laboratory for GeoInformation Science, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: yeeleung@cuhk.edu.hk.*
- *J.-H. Ma and W.-X. Zhang are with the Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, ROC. E-mail: jhmath@china.com.*

Manuscript received 4 May 1999; revised 10 May 2000; accepted 5 Sept. 2000. Recommended for acceptance by T. Ishida.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 109750.

particular, a large data set does not behave as nicely as stipulated by these assumptions. It is very common that inliers are out-numbered by outliers, making many robust methods fail.

To overcome the above difficulties and to utilize the advantages of parametric regression models, we first need to understand a complicated data set. It is of utmost importance to treat such data sets correctly. One appropriate way may be to view a complicated data set as a mixture of many populations. Suppose that a data set can be described by a finite number of regression models. If we view each regression model as a "population," then the data set is a mixture of a finite number of such "populations." Thus, identification of these models becomes the problem of mixture modeling.

In the literature, mixture modeling (variously known as a form of clustering, intrinsic classification, or numerical taxonomy) is the modeling of a statistical distribution by a mixture of other distributions, known as components or classes. Finite mixture densities have served as important models for the analysis of complex phenomena throughout the history of statistics [23]. This model deals with the unsupervised discovery of clusters within data [22]. In particular, mixtures of normal populations are most frequently studied and most widely employed in practice. In estimating mixture parameters, the maximum-likelihood (ML) method, in particular the maximum-likelihood estimator (MLE), has become the most extensively adopted approach [27]. Although the use of the expectation maximization (EM) algorithm greatly reduces the computational difficulty for MLE of mixture models, the EM algorithm still has drawbacks. The intolerably slow convergence of sequences of iterates generated by it in some applications is a typical example. Other methods such as the method of moments and the moment generating function (MGF) method generally involve the problem of simultaneously estimating all of the mixture parameters. Clearly, it is a very difficult task of estimation in large data sets.

As a certain type of mixtures of normal densities, a special mixture of regression models has been studied as a switching regression model in which observations are given on a random variable y and on a vector of fixed carriers x, and

$$Y = \begin{cases} X^T\beta_1 + e_1, & \text{with probability} \quad \lambda, \\ X^T\beta_2 + e_2, & \text{with probability} \quad 1-\lambda, \end{cases}$$

where $e_1 \sim N(0, \sigma_1^2), e_2 \sim N(0, \sigma_2^2)$, and λ as well as the vectors β_1, β_2 and σ_1, σ_2 are unknown. Switching regression models have many applications in practice (e.g., a wide class of industrial control problems [1], marketing research [21], economics [26], and fisheries research [13]). Lately, a potential application is found in the approximation of a nonlinear system by decomposing fuzzily the whole input space into several partial spaces and representing each input/output space with each linear equation [20]. Similar to the estimation of mixture parameters, it is still difficult to estimate the parameters of switching regression model in large data sets. The EM algorithm and its variants for this model is usually studied (e.g., [26], [3], [32]). Recently, a family of fuzzy C-regression models (FCRM) is used to study this model [13].

In addition to the effectiveness of an estimation method, another important feature that needs to be addressed is robustness. To be useful in practice, a method needs to be very robust, especially for large data sets. It means that the performance of a method should not be affected significantly by small deviations from the assumed model and it should not deteriorate drastically due to noise and outliers. Discussions on and comparison with several popular clustering methods from the point of view of robustness are summarized in [8]. Obviously, robustness in KDD is also absolutely necessary. Some attempts have been made in recent years (e.g., [18], [6]) and the problem needs to be further studied.

The purpose of this paper is to propose an effective and robust method for the mining of regression classes in large data sets, especially under contamination with noise. We first introduce a new concept named "regression-class" which is defined by a regression model. The concept is different from the existing conceptualization of class (cluster) based on common sense or a certain distance measure. As a generalization of classes, a regression class contains more useful information. The aforementioned switching regression models, for example, can be viewed as two regression classes. We assume that there is a finite number of this kind of regression classes in a large data set. Instead of considering the whole data set, sampling is used to identify the corresponding regression classes. A novel framework, formulated in a recursive paradigm, for mining multiple regression classes in a data set is then proposed. Based on a highly robust model-fitting (MF) estimator and an effective Gaussian mixture decomposition algorithm (GMDD) in computer vision [33], [34], the proposed method, coined regression-class mixture decomposition (RCMD), only involves the parameters of a regression class at each time of the mining process. Thus, it greatly reduces the difficulty of parametric estimation and achieves a high degree of robustness. We demonstrate that the RCMD method is suitable for small, medium, and large data sets and has many promising applications in a variety of disciplines including computer vision, pattern recognition, and economics.

It is necessary to point out that identifying some regression classes, such as "switching regressions," is different from the conventional classification problem, which is concerned with modeling the conditional distribution of Y given X. It also differs from other models, such as piecewise regression and regression tree, in which different subsets of X follow different regression models. The RCMD method not only can solve the identity problem of regression classes, but may also be extended to other models, such as piecewise regression.

In Section 2, we first define a regression class and investigate the effect of outliers on the ML estimator of mixture parameters. In Section 3, we propose the RCMD method and two associated algorithms for implementation. Some issues about the proposed method are also discussed. To substantiate the theoretical analysis, simulation runs and a real-life application are performed in Section 4 to evaluate the effectiveness and robustness of the proposed method. We then conclude the paper with a summary and plausible directions for further research.

2 Regression Classes and the Effect of Outliers on Their Identification

We propose in this section the concept of a "regression class" (abbreviated as "reg-class") and study the effect of outliers on their identification. Intuitively, a reg-class is equated with a regression model. To state it formally, for a fixed integer i, a reg-class G_i is defined by the following regression model with random carriers:

$$G_i : Y = f_i(X, \beta_i) + e_i, \tag{1}$$

where $Y \in R$ is the response variable, the explanatory variable that consists of carriers or regressors $X \in R^p$ is a random (column) vector with a probability density function (p.d.f.) $p(\bullet)$, the error term e_i is a random variable with a p.d.f. $\psi(u; \sigma_i)$ having a parameter σ_i, $Ee_i = 0$, and X and e_i are independent. Here, $f_i(\cdot, \cdot) : R^p \times R^{q_i} \to R$ is a known regression function, and $\beta_i \in R^{q_i}$ is an unknown regression parameter (column) vector. Although the dimension of β_i, q_i, may be different for different G_i, we usually take $q_i \equiv q$ for simplicity. Henceforth, we assume that e_i is distributed according to a normal distribution, i.e.,

$$\psi(u; \sigma_i) = \frac{1}{\sigma_i}\phi\left(\frac{u}{\sigma_i}\right), \tag{2}$$

where $\phi(\cdot)$ is the standard normal p.d.f.

For convenience of discussion, let

$$r_i(x, y; \beta_i) \equiv y - f_i(x, \beta_i). \tag{3}$$

Definition 1. *A random vector (X, Y) belongs to a* regression class *G_i (denoted as $(X, Y) \in G_i$) if it is distributed according to the regression model G_i.*

Thus, under Definition 1, a random vector $(X, Y) \in G_i$ implies that (X, Y) has a p.d.f.

$$p_i(x, y; \theta_i) = p(x)\psi(r_i(x, y; \beta_i); \sigma_i), \theta_i = (\beta_i^T, \sigma_i)^T. \tag{4}$$

For practical purpose, the following definition associated with Definition 1 often may be used.

Definition 2. *A data point (x, y) belongs to a* regression class *G_i (denoted as $(x, y) \in G_i$) if it satisfies $p_i(x, y; \theta_i) \geq b_i$, i.e.,*

$$G_i \equiv G_i(\theta_i) \equiv \{(x, y) : p_i(x, y; \theta_i) \geq b_i\}, \tag{5}$$

where the constant $b_i > 0$ is determined by

$$P[p_i(X, Y; \theta_i) \geq b_i] = a,$$

a is a probability threshold specified a priori *and approaches to one.*

For simplicity, we assume that there are m reg-classes $G_1, G_2, \ldots, G_m$ in a data set under study and that m is known in advance (indeed m can be determined at the end of the mining process when all plausible reg-classes have been identified). Our goal is to find all m reg-classes, to identify the parameter vectors and to make predication or interpretation by the models. To avoid large amount of computation, we need to randomly sample from a data set to search for the reg-classes. It is assumed for the present study that $(x_1, y_1), \ldots, (x_n, y_n)$ are the observed values of a random sample of size n taken from a data set. Thus, they can be considered as realized values of n independently and identically distributed (i.i.d.) random vectors with a common mixture distribution population

$$p(x, y; \theta) = \sum_{i=1}^{m} \pi_i p_i(x, y; \theta_i), \tag{6}$$

that is, they consist of random observations from m reg-classes with prior probabilities $\pi_1, \ldots, \pi_m (\pi_1 + \cdots + \pi_m = 1, \pi_i \geq 0, 1 \leq i \leq m), \theta^T = (\theta_1^T, \ldots, \theta_m^T)$.

In what follows, we first investigate theoretically the issues of the search for reg-classes in the presence of noise contamination.

We first consider the case in which $\pi_1, \ldots, \pi_m$ are known. In this case, all unknown parameters consist of the aggregate vector $\theta = (\theta_1^T, \ldots, \theta_m^T)^T$. If the vector $\theta_0^T = (\theta_1^{0^T}, \ldots, \theta_m^{0^T})$ of true parameters is known a priori, and the outliers are absent ($\varepsilon_i \equiv 0, 1 \leq i \leq m$), then the posterior probability that (x_j, y_j) belongs to G_i is given by

$$\tau_i(x_j, y_j; \theta_i^0) = \frac{\pi_i p_i(x_j, y_j; \theta_i^0)}{\sum_{k=1}^{m} \pi_k p_k(x_j, y_j; \theta_k^0)}, 1 \leq i \leq m.$$

A partitioning of the sample $Z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ into m reg-classes can be made by assigning each (x_j, y_j) to the population to which it has the highest estimated posterior probability of belonging to G_i if

$$\tau_i(x_j, y_j; \theta_i^0) > \tau_k(x_j, y_j; \theta_k^0), 1 \leq k \leq m, k \neq i.$$

This is just the Bayesian decision rule:

$$\begin{aligned} d &= d(x, y; \theta_0) \\ &= \arg \max_{1 \leq i \leq m} \left[\pi_i p_i(x, y; \theta_i^0)\right], x \in R^p, y \in R, 1 \leq d \leq m, \end{aligned} \tag{7}$$

which classifies the sample Z and "new" observation with minimal error probability. As θ_0 is unknown, the so-called "plug-in" decision rule is often used:

$$d = d(x, y; \hat{\theta}_0) = \arg \max_{1 \leq i \leq m} [\pi_i p_i(x, y; \hat{\theta}_i^0)], \tag{8}$$

where $\hat{\theta}_0$ is the MLE of θ_0 constructed by the sample Z from the mixture population,

$$\hat{\theta}_0 = \arg \max_{\theta \in \Theta} l(\theta), \tag{9}$$

$$l(\theta) = \ln \prod_{j=1}^{n} p(x_j, y_j; \theta) = \sum_{j=1}^{n} \ln p(x_j, y_j; \theta), \tag{10}$$

where Θ is a parameter space.

We consider the case in which $p_i(x, y; \theta_i)$ is contaminated, i.e., the ε_i-contaminated neighborhood is:

$$\begin{aligned} U(\varepsilon_i) = \big\{ & p_i^{\varepsilon}(x, y; \theta_i) : p_i^{\varepsilon}(x, y; \theta_i) \\ & = (1 - \varepsilon_i) p_i(x, y; \theta_i) + \varepsilon_i h_i(x, y) \big\}, \end{aligned}$$

where $h_i(x, y)$ is any p.d.f. of outliers in G_i, ε_i is an unknown fraction of an outlier present in G_i.

Now, we investigate the effect of outliers on the MLE $\hat{\theta}_0$ under ε-contaminated models. In this case, Z is the random sample from the mixture p.d.f.:

$$p_\varepsilon(x,y;\theta_0)=\sum_{i=1}^{m}\pi_i p_i^\varepsilon(x,y;\theta_i^0). \tag{11}$$

Let ∇_θ^k be the operator of the kth order differentiation with respect to θ, $\mathbf{0}$ be a zero matrix with all elements being zero and $\mathbf{1}$ be a matrix with all elements being 1. Denote $p_0(x,y;\theta)=p(x,y;\theta)$,

$$\begin{aligned} I_\varepsilon(\theta;\theta_0) &= -E_\varepsilon[\ln p_0(X,Y;\theta)] \\ &= -\iint_{R^{p+1}} \ln p_0(x,y;\theta)p_\varepsilon(x,y;\theta_0)dxdy, \end{aligned} \tag{12}$$

$$B_i(\theta)=\iint_{R^{p+1}}[h_i(x,y)-p_i(x,y;\theta_i^0)]\ln p_0(x,y;\theta)dxdy, \tag{13}$$

$$J_\varepsilon(\theta_0)=-\iint_{R^{p+1}} p_\varepsilon(x,y;\theta_0)\nabla_\theta^2 \ln p_0(x,y;\theta)|_{\theta=\theta_0}dxdy. \tag{14}$$

It can be observed that $I_0(\theta_0;\theta_0)$ is the Shannon entropy for the hypothetical mixture $p_0(x,y;\theta)$. Furthermore, a simple calculation shows that $J_0(\theta_0)$ is the Fisher information matrix

$$J_0(\theta_0)= \iint_{R^{p+1}} p_0(x,y;\theta_0)\nabla_\theta \ln p_0(x,y;\theta)[\nabla_\theta \ln p_0(x,y;\theta)]^T|_{\theta=\theta_0}dxdy,$$

and in regularity conditions,

$$\nabla_\theta I_0(\theta;\theta_0)|_{\theta=\theta_0}=\mathbf{0}, \nabla_\theta^2 I_\varepsilon(\theta;\theta_0)|_{\theta=\theta_0}=J_\varepsilon(\theta_0). \tag{15}$$

Theorem 1. *If the family of p.d.f. $p(x,y;\theta)$ satisfies the regularity condition [19], the function $I_0(\theta;\theta_0)$, $B_i(\theta)$ are thrice differentiable with respect to $\theta\in\Theta$, and the point $\theta_\varepsilon=\arg\min_{\theta\in\Theta} I_\varepsilon(\theta;\theta_0)$ is unique, then the MLE $\hat{\theta}$ under ε-contamination is almost surely convergent, i.e.,*

$$\hat{\theta}\xrightarrow{a.s.}\theta_\varepsilon(n\to\infty) \tag{16}$$

and $\theta_\varepsilon\in\Theta$ satisfies the asymptotic expansion:

$$\theta_\varepsilon=\theta_0+[J_\varepsilon(\theta_0)]^{-1}\sum_{i=1}^{m}\varepsilon_i\pi_i\nabla_\theta B_i(\theta_0)+\mathrm{O}\left(||\theta_\varepsilon-\theta_0||^2\right)\mathbf{1} \tag{17}$$

(see Appendix for the proof).

Remark 1. It can be observed from Theorem 1 that in the presence of outliers in the sample, the estimator $\hat{\theta}$ can become inconsistent (see (16), (17)). It should be noted that $|\nabla_\theta B_i(\theta)|$ depends on the contaminating density $h_i(x,y)$ and may have sufficiently large value $(1\le i\le m)$.

Corollary 1. *In the setting of Theorem 1, $\hat{\theta}$ has an influence function*

$$IF(x,y;\hat{\theta})=[J_0(\theta_0)]^{-1}\nabla_\theta \ln p_0(x,y;\theta)|_{\theta=\theta_0} \tag{18}$$

(see Appendix for the proof).

Remark 2. The influence function (IF) is an important concept in robust statistics. It can measure the effect of an additional observation in any point (x,y) on the estimator $\hat{\theta}$.

Now, we discuss briefly the case in which $\pi_1,\ldots,\pi_m$ are unknown. Let $\pi=(\pi_1,\ldots,\pi_m)^T$, $\varphi=(\pi^T,\theta^T)^T$, $L(\varphi)=\ln\prod_{j=1}^{n}p_\varepsilon(x_j,y_j;\theta)$, and

$$\eta_i(x_j,y_j;\varphi)=\frac{\pi_i p_i^\varepsilon(x_j,y_j;\theta_i)}{\sum_{k=1}^{m}\pi_k p_k^\varepsilon(x_j,y_j;\theta_k)}, 1\le i\le m.$$

By the method in [23], we can find that the MLE of $\varphi, \hat{\varphi}=(\hat{\pi}^T,\hat{\theta}^T)^T$, satisfies

$$\nabla_{\pi_k}L(\varphi)=\sum_{j=1}^{n}\left(\frac{\eta_k(x_j,y_j;\theta_k)}{\pi_k}-\frac{\eta_m(x_j,y_j;\theta_m)}{\pi_m}\right)=0,$$

$$\nabla_{\theta_k}L(\varphi)\Big|_{\theta_k=\hat{\theta}_k}= \sum_{j=1}^{n}\eta_k(x_j,y_j;\hat{\varphi})\nabla_\theta \ln p_k^\varepsilon(x_j,y_j;\theta_k)\Big|_{\theta_k=\hat{\theta}_k}=\mathbf{0},\ 1\le k\le m.$$

It can be observed that it may be very difficult to get $\hat{\varphi}$ when m is large, because too many parameters are involved. As a matter of fact, the ML method for directly estimating the parameters of mixture densities actually has many practical implementation difficulties [31], [33].

In the next section, we propose an effective and feasible method which involves relatively few parameters in the mining of reg-classes.

3 The Regression-Class Mixture Decomposition (RCMD) Method

Based on the GMDD algorithm [33] and the MF estimator [34], we propose in this section the RCMD method for mining multiple reg-classes in large data set. To a certain extent, the method can be viewed as an extension of GMDD and MF.

3.1 The RCMD Estimator

Developing methods to resist the effect of outliers is an aim of robust statistics. It is well-known that almost all of the robust methods tolerate only less than 50 percent of outliers. When there are multiple reg-classes in a data set, they cannot offer a solution to the identification of these classes, because it is very common that the proportion of outliers with respect to a single class is more than 50 percent. Recently, several more robust methods have been developed for computer vision. For example, MINPRAN [29] is perhaps the first technique that reliably tolerates more than 50 percent of outliers without assuming a known bound for inliers. The method assumes that the outliers are randomly distributed within the dynamic range of the sensor, and the noise (outlier) distribution is known. The assumptions of MINPRAN restrict its generality in practice.

Another highly robust estimator is the MF estimator [34], which is developed for a simple regression problem without carriers. It does not need assumptions such as those in MINPRAN. Indeed, no requirement is imposed on the distribution of outliers. So, it seems to be more applicable to a complex data set. Extended on the ideas of

the MF estimator and GMDD, we now derive the RCMD estimator in the analysis to follow.

With respect to a particular density or structure, all other densities or structures in a mixture can be readily classified as part of the outlier category in the sense that these other observations obey different statistics. Thus, a mixture density can be viewed as a contaminated density with respect to each cluster in the mixture. When all of the observations for a single density are grouped together, the remaining observations (clusters and true outliers) can then be considered to form an unknown outlier density. According to this idea, the mixture p.d.f. in (11) with respect to G_i can be rewritten as

$$\begin{aligned} &p_\varepsilon(x,y;\theta)\\ &\quad= \pi_i(1-\varepsilon_i)p_i(x,y;\theta_i)+\pi_i\varepsilon_i h_i(x,y)+\sum_{j\neq i}^{m}\pi_j p_j^{\varepsilon}(x,y;\theta_j)\\ &\quad\equiv \pi_i(1-\varepsilon_i)p_i(x,y;\theta_i)+[1-\pi_i(1-\varepsilon_i)]g_i(x,y). \end{aligned} \tag{19}$$

Ideally, a sample point (x_k, y_k) from the above mixture p.d.f. is classified as an *inlier* if it is realized from $p_i(x,y;\theta_i)$ or as an outlier otherwise (i.e., it comes from the p.d.f. $g_i(x,y)$).

Now, the given data set $Z=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ is generated by the mixture p.d.f. $p_\varepsilon(x,y;\theta)$, i.e., it comes from $p_i(x,y;\theta_i)$ with probability $\pi_i(1-\varepsilon_i)$ together with an unknown outlier $g_i(x,y)$ with probability $[1-\pi_i(1-\varepsilon_i)]$.

Let D_i be the subset of all inliers with respect to G_i and $\bar{D}_i$ be its complement. From the Bayesian classification rule, we have

$$D_i=\left\{(x_j,y_j):p_i(x_j,y_j;\theta_i)>\tfrac{1-\pi_i+\pi_i\varepsilon_i}{\pi_i(1-\varepsilon_i)}g_i(x_j,y_j)\right\},\ \ \bar{D}_i=Z-D_i. \tag{20}$$

Define

$$\begin{aligned} d_i^0&=\min\{p_i(x_j,y_j;\theta_i):(x_j,y_j)\in D_i\},\\ d_i^1&=\max\{p_i(x_j,y_j;\theta_i):(x_j,y_j)\in \bar{D}_i\}. \end{aligned}$$

Ideally, the likelihood of any inlier being generated by $p_i(x,y;\theta_i)$ is greater than the likelihood of any outlier being generated by $g_i(x,y)$. Hence, it is argued in [33] that we may assume that $d_i^0>d_i^1$. Therefore, the Bayesian classification becomes

$$D_i=\left\{(x_j,y_j):p_i(x_j,y_j;\theta_i)>\frac{1-\pi_i+\pi_i\varepsilon_i}{\pi_i(1-\varepsilon_i)}\delta_i\right\}, \tag{21}$$

where we can choose

$$\delta_i\in[\pi_i(1-\varepsilon_i)d_i^1/(1-\pi_i+\pi_i\varepsilon_i),\pi_i(1-\varepsilon_i)d_i^0/(1-\pi_i+\pi_i\varepsilon_i)].$$

This implies that if we assume that

$$g_i(x_1,y_1)=\ldots=g_i(x_n,y_n)=\delta_i,$$

then, we would get equivalent results. Using this assumption, (19) becomes

$$p_\varepsilon(x,y;\theta)=\pi_i(1-\varepsilon_i)p_i(x,y;\theta_i)+(1-\pi_i+\pi_i\varepsilon_i)\delta_i.$$

The log-likelihood function of observing Z corresponding to (10) under ϵ-contamination becomes

$$\begin{aligned} &l(\theta_i)=\\ &n\ln[\pi_i(1-\varepsilon_i)]+\sum_{j=1}^{n}\ln\left[p_i(x_j,y_j;\theta_i)+\frac{1-\pi_i+\pi_i\varepsilon_i}{\pi_i(1-\varepsilon_i)}\delta_i\right]. \end{aligned}$$

Thus, in order to estimate θ_i from Z, we need to maximize $l(\theta_i)$ with each δ_i subject to $\sigma_i>0$. Since the maximization of $l(\theta_i)$ at δ_i with respect to θ_i is equivalent to maximizing the G_i model-fitting function

$$l_i(\theta_i;t_i)\equiv\sum_{j=1}^{n}\ln[p_i(x_j,y_j;\theta_i)+t_i] \tag{22}$$

at t_i with respect to θ_i, provided that $t_i=(1-\pi_i+\pi_i\varepsilon_i)\delta_i/[\pi_i(1-\varepsilon_i)]$, then we can discuss the problem of maximizing $l(\theta_i)$ subject to $\sigma_i>0$. We henceforth shall refer to each "t_i" (≥ 0) as a partial model [33]. Since each t_i corresponds to a value δ_i of outlier distribution $g_i(x,y)$, we only use the partial information about the model without the knowledge of the whole shape of $g_i(x,y)$.

Definition 3. *For a reg-class G_i and the data set $Z=\{(x_1,y_1),\cdots,(x_n,y_n)\}$, the* t-level set *of G_i is defined as*

$$G_i(\theta_i;t)=\{(x_j,y_j):p_i(x_j,y_j;\theta_i)>t\}, \tag{23}$$

the t-level support set *of an estimator $\hat{\theta}_i$ for θ_i is defined as $G_i(\hat{\theta}_i;t)$.*

Remark 3. According to this concept, $G_i(\theta_i;t)$ is the subset of all inliers with respect to G_i at a partial model t. Maximizing (22) may be approximately interpreted as maximizing "likelihood" over the t-level set of G_i. It should be noted that the capacity of $G_i(\theta_i;t)$ will decrease as a partial model level t increases. Moreover, the t-level support set of an estimator $\hat{\theta}_i$ reflects the extent to which the data set supports this estimator at partial model level t.

Definition 4. *The RCMD estimator of the parametric vector θ_i for a reg-class G_i is defined by*

$$\hat{\theta}_i^t=\arg\max_{\theta_i} l_i(\theta_i;t_i),\quad \theta_i=(\beta_i^T,\sigma_i)^T,\quad \sigma_i>0.$$

When $m=1$ and the random carriers disappear in (1), the RCMD estimator becomes an univariate MF estimator. In particular, when X is distributed uniformly (i.e., $p(x)\equiv$ constant in some domain) and $e_i\sim N(0,\sigma_i^2)$, the maximization of $l_i(\theta_i;t_i)$ is equivalent to maximizing

$$\bar{l}_i(\theta_i;\bar{t}_i)\equiv\sum_{j=1}^{n}\ln\big(\psi[y_j-f_i(x_j,\beta_i);\sigma_i]+\bar{t}_i\big),$$

where $\bar{t}_i=t_i/c$. For simplicity and without further notification, we will still denote $\bar{t}_i$ and $\bar{l}_i$ by t_i and l_i, respectively. That is, the above expression is rewritten as

$$l_i(\theta_i;t_i)\equiv\sum_{j=1}^{n}\ln\big(\psi[y_j-f_i(x_j,\beta_i);\sigma_i]+t_i\big). \tag{24}$$

In this case, the corresponding expressions in (23) and (5) become respectively

$$G_i(\theta_i;t_i)=\{(x_j,y_j):\psi[r_i(x_j,y_j;\beta_i);\sigma_i]>t_i\}, \tag{25}$$

$$G_i(\theta_i) = \{(x, y) : |r_i(x, y; \beta_i)| \le 3\sigma_i\}, \quad (26)$$

which is based on the three σ-criterion of the normal distribution (i.e., a in (5) is 0.9972).

Theorem 2. *Under the conditions of Theorem 1, the almost sure convergence of $\hat{\theta}_i^t$ which maximizes $l_i(\theta_i; t_i)$ in (22) holds:*

$$\hat{\theta}_i^t \xrightarrow{a.s.} \theta_{i,\varepsilon}^t,$$

and an asymptotic expression similar to (17) holds also, where $\theta_{i,\varepsilon}^t = \arg\min_{\theta_i \in \Theta_i} I_\varepsilon^t(\theta_i; \theta_o)$,

$$I_\varepsilon^t(\theta_i; \theta_0) = -E_\varepsilon \ln[p_i(X, Y; \theta_i) + t_i],$$

and $B_i(\theta)$ in (17) becomes

$$B_i^t(\theta_i) = \iint_{R^{p+1}} [h_i(x, y) - p_i(x, y; \theta_i^0)] \ln[p_i(x, y; \theta_i) + t_i] dx\, dy.$$

(The proof is parallel to that of Theorem 1 and is omitted here.)

Remark 4. It is clear from Theorem 2 that $\hat{\theta}_i^t$ maximizing $l_i(\theta_i; t_i)$ may be a biased estimator. However, we can revise it by an unbiased LS estimator afterwards.

Now, we state the RCMD method in brief as follows (similar to [33]):

At each selected partial model $t_i^{(s)}$, $s = 0, 1, \ldots, S$, we maximize $l_i(\theta_i; t_i^{(s)})$ with respect to β_i and σ_i by using an iterative algorithm beginning with a randomly chosen initial $\beta_i^{(0)}$ or by using a genetic algorithm (GA). Having solved $\max_{\beta_i, \sigma_i} l_i(\theta_i; t_i^{(s)})$ for $\hat{\beta}_i(t_i^{(s)})$ and $\hat{\sigma}_i(t_i^{(s)})$, we calculate the possible reg-class $G_i(\hat{\theta}_i(t_i^{(s)}))$ followed by the test of normality on $G_i(\hat{\theta}_i(t_i^{(s)}))$. If the test statistic is not significant (usually at level $\alpha = 0.01$), then the hypothesis that the respective distribution is normal should be accepted and a valid reg-class, $G_i(\hat{\theta}_i(t_i^{(s)}))$, has been determined, otherwise we proceed to the next partial model if the upper bound $t_i^{(S)}$ has not been reached. It may be said that the identity of each $G_i(\hat{\theta}_i(t_i^{(s)}))$ is based on its *t*-level set.

After a valid reg-class has been detected, it is extracted from the current data set, and the next reg-class will be identified in the new size-reduced data set. Individual reg-classes continue to be estimated recursively until there are no more valid reg-classes, or the size of the new data set gets to be too small for estimation. Thus, the proposed method can handle an arbitrary number of reg-class models with single reg-class extraction. The flowchart of the RCMD method is depicted in Fig. 1.

Since the Kolmogorov-Smirnov (K-S) D test is only valid when the mean and standard deviation of the normal distribution are known a priori and not estimated from data, alternative tests may then be considered. In recent years, some preferred tests of normality have been proposed. When the sample contains at most up to 2,000 observations, we can select the Shapiro-Wilks' (S-W) [28] W test owing to its good power properties as compared to a wide range of alternative tests and we do not need to estimate parameters. Otherwise, the K-S test should be used.

In particular, when $q_i = 1$ and $f_i(X, \beta_i) = \beta_i$ in (1), the RCMD method becomes the GMDD algorithm for the univariate case.

3.2 The Computational Formulas

3.2.1 Using an Iterative Algorithm

The gradient ascent rule is used to solve each maximization of (22) for a specified t_i under the normal error distribution in (2). The gradients $\nabla_{\beta_i} l_i(\theta_i)$ and $\nabla_{\sigma_i} l_i(\theta_i)$ can be derived as

$$\begin{aligned}
&\nabla_{\beta_i} l_i(\theta_i; t_i) \\
&\quad = \nabla_{\beta_i} \sum_{j=1}^{n} \ln[p_i(x_j, y_j; \theta_i) + t_i] \\
&\quad = \sum_{j=1}^{n} \frac{1}{p_i(x_j, y_j; \theta_i) + t_i} \nabla_{\beta_i} [p_i(x_j, y_j; \theta_i)] \\
&\quad = \sum_{j=1}^{n} \frac{p_i(x_j, y_j; \theta_i)}{p_i(x_j, y_j; \theta_i) + t_i} \nabla_{\beta_i} [\ln p_i(x_j, y_j; \theta_i)] \\
&\quad = \sum_{j=1}^{n} \lambda_{ij} \nabla_{\beta_i} [\ln p_i(x_j, y_j; \theta_i)] \\
&\quad = \sum_{j=1}^{n} \lambda_{ij} \left[\nabla_{\beta_i} \ln p(x_j) + \nabla_{\beta_i} \ln\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right) - \nabla_{\beta_i} \left(\frac{r_{ij}^2}{2\sigma_i^2}\right) \right] \\
&\quad = \sigma_i^{-2} \sum_{j=1}^{n} \lambda_{ij} r_{ij} \nabla_{\beta_i} f_i(x_j, \beta_i), \\
&\nabla_{\sigma_i} l_i(\theta_i; t_i) \\
&\quad = \sum_{j=1}^{n} \lambda_{ij} \nabla_{\sigma_i} \ln\big[p_i(x_j, y_j; \theta_i)\big] = \sum_{j=1}^{n} \nabla_{\sigma_i} \left[\ln\left(\frac{1}{\sigma_i}\right) \phi\left(\frac{r_{ij}}{\sigma_i}\right) \right] \\
&\quad = \sum_{j=1}^{n} \nabla_{\sigma_i} \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right) - \frac{r_{ij}^2}{2\sigma_i^2} \right] = \sigma_i^{-3} \sum_{j=1}^{n} \lambda_{ij} (r_{ij}^2 - \sigma_i^2),
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{ij} &\equiv p_i(x_j, y_j; \theta_i) / [p_i(x_j, y_j; \theta_i) + t_i], \\
r_{ij} &\equiv r_i(x_j, y_j; \beta_i) = y_j - f_i(x_j, \beta_i).
\end{aligned}$$

To maximize $l_i(\theta_i; t_i)$ with respect to θ_i for each $t_i \ge 0$ under (2), the following stationary equation must be satisfied:

$$\sum_{j=1}^{n} \lambda_{ij} r_{ij} \nabla_{\beta_i} f_i(x_j, \beta_i) = 0, \quad (27)$$

$$\sum_{j=1}^{n} \lambda_{ij} (r_{ij}^2 - \sigma_i^2) = 0. \quad (28)$$

In particular, when $f_i(x, \beta_i) = x^T \beta_i$, the equation in (27) becomes

$$\sum_{j=1}^{n} \lambda_{ij} r_{ij} x_j = \sum_{j=1}^{n} \lambda_{ij} y_j x_j - \sum_{j=1}^{n} \lambda_{ij} x_j x_j^T \beta_i = 0. \quad (29)$$

If $\sum_{j=1}^{n} \lambda_{ij} x_j x_j^T$ is nonsingular, then we can obtain a nominal expressions for $\hat{\theta}_i = (\hat{\beta}_i^T, \hat{\sigma}_i)^T$ which maximizes $l_i(\theta_i; t_i)$:

$$\hat{\beta}_i = \left[\sum_{j=1}^{n} \hat{\lambda}_{ij} x_j x_j^T \right]^{-1} \sum_{j=1}^{n} \hat{\lambda}_{ij} y_j x_j, \quad (30)$$

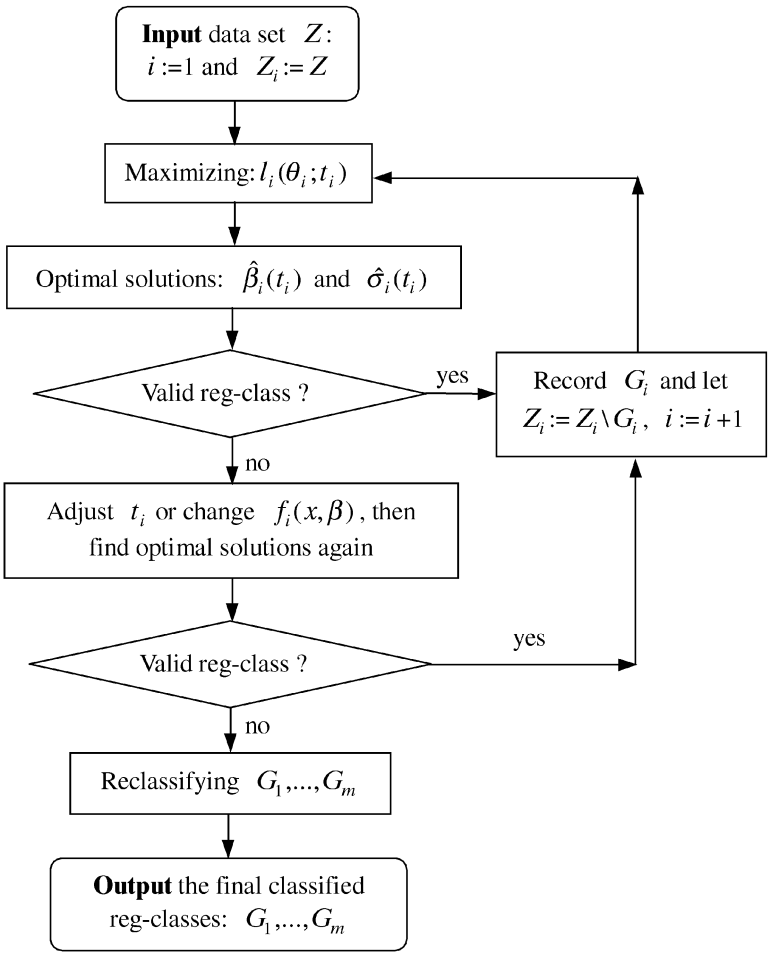

Fig. 1. Flowchart of the RCMD method.

$$\hat{\sigma}_i^2 = \left[\sum_{j=1}^{n} \hat{\lambda}_{ij}\right]^{-1} \sum_{j=1}^{n} \hat{\lambda}_{ij} \hat{r}_{ij}^2, \tag{31}$$

where $\hat{\lambda}_{ij} \equiv \hat{p}_{ij}/[\hat{p}_{ij} + t_i]$, $\hat{p}_{ij} \equiv p_i(x_j, y_j; \hat{\theta}_i)$,

$$\hat{r}_{ij} \equiv r_i(x_j, y_j; \hat{\beta}_i) = y_j - x_j^T \hat{\beta}_i.$$

It should be noted that the right hand sides of the equations in (30) and (31) still contain $\hat{\beta}_i$ and $\hat{\sigma}_i^2$.

When the data set is of small or medium size and a good starting value can be given, results obtained by the iterative algorithm are satisfactory. For large data sets or when difficulty in obtaining a good starting value is encountered, the iterative algorithm whose performance depends on the choice of starting values may not converge to a true solution or converge very slowly, or even does not converge. We propose another algorithm in the following section to accomplish the task.

3.2.2 Using a Genetic Algorithm

Genetic algorithm (GAs) [25] were inspired by biological evolutions. They, in brief, carry out a simulated form of evolution on populations of chromosomes representing subjects of interest. In addition to producing a more global search, the multi-point search makes GAs highly implementable in parallel machines. Thus, the RCMD method with GAs is suitable to mine multiple regression-classes in a large data set. GAs use chance efficiently in their exploitation of prior knowledge to rapidly locate near-optimal solutions. If we have some domain specific knowledge on the ranges of the optimal parameters, we should use it in order to reduce the search space and give a good initialization.

Although some GAs have become quite complex, good results can be achieved with relatively simple GAs. The simple GA used in this study consists of the reproduction, crossover, and mutation operators basic to the canonical GAs. The fitness function is $l_i(\theta_i; t_i)$ in (22). We will use θ_i as a chromosome to represent the parametric estimation, define a P_{size} as the number of chromosomes, and initialize the chromosomes randomly. A biased *roulette wheel* is used as a simple implementation of the selection operator. For the crossover operator, we use single-point crossover with probability P_c. For the mutation operator, we perform random alteration of the value of a string position with probability P_m.

3.3 The Computational Steps

For each selected partial model "t_i," the RCMD method using the GA procedure can be implemented via the following steps:

Step 1. Set the population size P_{size}, P_c, and P_m. For the fitness function, in (22) find an optimal solution $\hat{\theta}_i = (\hat{\beta}_i^T, \hat{\sigma}_i)^T$.

Step 2. Once the optimal $\hat{\beta}_i$ and $\hat{\sigma}_i$ are obtained, perform the S-W (or K-S) test of normality for each $G_i(\hat{\theta}_i(t_i))$. If the W (or D) statistic is significant, then the hypothesis that the error distribution of reg-classes is normal should be rejected, set another t_i and go to **Step 1**. Otherwise, the estimators $\hat{\beta}_i$ and $\hat{\sigma}_i$ will be obtained, and a valid reg-class $G_i(\hat{\theta}_i(t_i))$ will be determined and removed from the data set, go to **Step 1** and start to find another reg-class.

Step 3. The final precise reg-classes are formed by applying the least-squares (LS) method with respect to all of the points falling into $G_i(\hat{\theta}_i(t_i))$.

It is well-known that the LS method can provide the most effective and unbiased estimators in the normal case without outliers, and its computation is relatively simple. We can then select the LS estimation as the final result since $G_i(\hat{\theta}_i(t_i))$ does not contain outliers with respect to G_i.

For the iterative algorithm, **Step 1** is replaced by **Step 1′** and other steps remain unchanged.

Step1′.

1. Set $k := 0$. Randomly choose $\hat{\beta}_i^{(0)}$, then calculate

$$\left[\hat{\sigma}_i^{(0)}\right]^2 = \frac{1}{n-p}\sum_{j=1}^{n}\left[y_j - x_j^T\hat{\beta}_i^{(0)}\right]^2,$$

$$\hat{\lambda}_{ij}^{(0)} = p_i\left(x_j, y_j; \hat{\theta}_i^{(0)}\right)\Big/\left[p_i\left(x_j, y_j; \hat{\theta}_i^{(0)}\right) + t_i\right] \text{ and}$$
$$\hat{r}_{ij}^{(0)} = y_j - x_j^T\hat{\beta}_i^{(0)}.$$

2. Set $k := k+1$. Given the kth estimates $\hat{\beta}_i^{(k)}$ and $\hat{\sigma}_i^{(k)}$ at the kth iterative step, $\hat{r}_{ij}^{(k)} = r_i(x_j, y_j; \hat{\beta}_i^{(k)})$, $\hat{p}_{ij}^{(k)} = p_i(x_j, y_j; \hat{\theta}_i^{(k)})$ and $\hat{\lambda}_{ij}^{(k)} = \hat{p}_{ij}^{(k)}/[\hat{p}_{ij}^{(k)} + t_i]$ are first calculated. Then, the (k+1)th estimates $\hat{\beta}_i^{(k+1)}$ and $\hat{\sigma}_i^{(k+1)}$ are obtained as

$$\hat{\beta}_i^{(k+1)} = \left[\sum_{j=1}^{n}\hat{\lambda}_{ij}^{(k)}x_j x_j^T\right]^{-1}\sum_{j=1}^{n}\hat{\lambda}_{ij}^{(k)}y_j x_j, \qquad (32)$$

$$\left[\hat{\sigma}_i^{(k+1)}\right]^2 = \left[\sum_{j=1}^{n}\hat{\lambda}_{ij}^{(k)}\right]^{-1}\sum_{j=1}^{n}\hat{\lambda}_{ij}^{(k)}\left[\hat{r}_{ij}^{(k)}\right]^2. \qquad (33)$$

3. If $||\hat{\beta}_i^{(k+1)} - \hat{\beta}_i^{(k)}|| < \delta_1$ and $|\hat{\sigma}_i^{(k+1)} - \hat{\sigma}_i^{(k)}| < \delta_2$, then go to **Step 2**, otherwise, go to 2.

To illustrate the effectiveness and robustness of the proposed method in the presence of data mixture, we first give a simple example ($p = 1$).

Example 1. Assuming that there are nine points in a data set, where five points fit the regression model: $Y = \beta_1 X + e_1$, $e_1 \sim N(0, \sigma_1^2)$, $\beta_1 = 1, \sigma_1 = 0.1$, and the others fit the regression model: $Y = \beta_2 X + e_2$, $e_2 \sim N(0, \sigma_2^2)$, $\beta_2 = 0, \sigma_2 = 0.1$ (see Fig. 2a). Now, we use these data points to identify the two regression classes. If we select $t_1 = 0.1$, the objective function is the G_1 model-fitting function

$$l_1(\theta_1; t_1) = \sum_{j=1}^{9}\ln\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_j - x_j\beta)^2}{2\sigma^2}\right) + 0.1\right].$$

This function has two obvious peaks, with each corresponds to a relevant reg-class. Using the iterative algorithm or genetic algorithm, the two reg-classes are easily identified. It is clearly shown in the contour plot of this function (Fig. 2c). For example, using the GA procedure, we can find: $\hat{\beta}_1 = 1.002, \hat{\sigma}_1 = 0.109$, and $l_{\max} = -2.167$. Using more exact maximization method, we obtain $\hat{\beta}_1 = 1.00231, \hat{\sigma}_1 = 0.109068$, and $l_{\max} = -2.16715$. The difference between the estimated values and the true parameters is in fact very small. On the other hand, if there is only one reg-class in this set (see Fig. 2b), our objective function is still very sensitive to this change. It can also find the only reg-class in the data set. There is only one peak which represents the reg-class (Fig. 2d).

3.4 Some Comments

3.4.1 About the Partial Models

From (22), it can be observed that maximizing $l_i(\theta_i; t_i)$ is equivalent to minimizing

$$\frac{1}{2\sigma_i^2}\sum_{j=1}^{n}[y_j - f_i(x_j, \beta_i)]^2 + n\ln\left(\sqrt{2\pi}\sigma_i\right) - \sum_{j=1}^{n}\ln p(x_j),$$

when $t_i = 0$. Obviously, the minimization of this expression with respect to $\theta_i = (\beta_i^T, \sigma_i)^T$ can be directly accomplished by the minimization with respect to β_i followed by σ_i, which results in the ordinary least squares (OLS) estimates of β_i. They are not robust and in the presence of outliers they give a poor estimation (see Fig. 5).

However, when $t_i > 0$, the situation is quite different. In fact, the parameter estimation with $t_i > 0$ is fairly robust and the estimated result can be greatly improved (see Fig. 5). The introduction to a partial model "$t_i > 0$" not only represents the consideration of outliers, but is also the simplification of this consideration in order to perform well. It is just the advantage of our method.

With Example 1, we can also show such a fact: the partial model t plays an important role in the mining of multiple reg-classes and if t is selected in a certain range, the maximization of objective function $l(\theta; t)$ is then meaningful. From (23), there is a range of t such that the t-level set is nonempty. In this range, reg-classes contained in the data set can be identified.

The experiment shows that for the data in Example 1, even when t is very small (10^{-3}), the RCMD method is still effective. However, it becomes invalid when t equals zero. When t changes from a very small positive number to approximately 5, the method remains valid. Once t exceeds 5, the greater is t, the more difficult it becomes for the RCMD method to identify the reg-classes. In general, the determination of the range of t depends on specific data set.

It has been shown that the model-fitting (MF) estimator is data-confusion resistant and is not adversely affected when the proportion of bad data increases ([34]). Due to the large range of choices available for a satisfactory partial

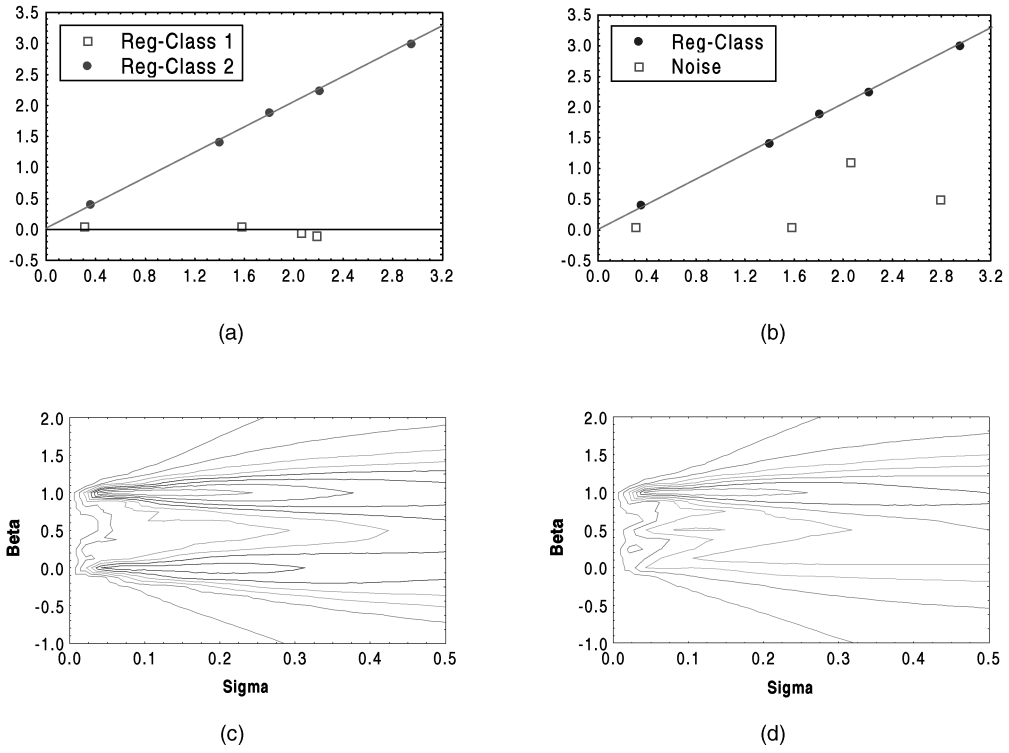

Fig. 2. Results obtained by the RCMD method for two reg-classes and one reg-class. (a) Scatterplot for two reg-classes. (b) Scatterplot for one reg-class. (c) Contour plot of objective function for (a). (d) Contour plot of objective function for (b).

model "t," it is proposed in [33] to integrate a 1D sequential search for the partial model t using a Monte Carlo random search method for different initializations of normal mean parameters. This search process also improves the speed of convergence and the efficiency in the search of true normal clusters. For a partial model t in the RCMD method, this search method still can be utilized.

When there is no significant overlapping of the reg-classes, the greater the given value of t_i is, the smaller the estimated value of σ_i becomes (see Fig. 5 in Example 2 to be discussed). This can be interpreted by the maximum $p(x)/(\sqrt{2\pi}\sigma_i)$ of $p_i(x, y; \theta_i)$: it is because the maximum of $l_i(\theta_i; t_i)$ will be achieved on the set $G_i(\theta_i; t_i)$, so $p(x)/(\sqrt{2\pi}\sigma_i)$ should at least exceed t_i. Nevertheless, it is not such a case when the reg-classes overlap significantly (see the case in Example 2). Experimentally, the choice of t_i has an influence on the estimation of σ_i, but almost has no influence on that of β_i.

3.4.2 On the Overlapping of Reg-Classes

When there is a significant overlapping of reg-classes, the assumption "with respect to each cluster the points belonging to other clusters can be considered as outliers," as well as the assumption in GMDD are difficult to justify. Removing one cluster may destroy the structures of other overlapping clusters which in turn may affect the end results of the structure mining process. Some effort has recently been devoted to solve this problem. The inlier classification procedure proposed in [7], for example, can be used to solve problems in which two models are intersecting, touching, or in the vicinity of each other.

For the overlapping of two reg-classes, we give here another data classification rule which is different from those in [7]. Once the parameters of two reg-classes G_i and G_j have been identified by the RCMD method, we can adopt the following rule for the assignment of data points in $G_i \cap G_j$: a data point $(x_k, y_k) \in G_i \cap G_j$ is assigned to G_i if

$$p_i(x_k, y_k; \hat{\theta}_i) > p_j(x_k, y_k; \hat{\theta}_j). \tag{34}$$

Combining (26) and (34), we can reclassify the data set into reg-classes. That is, although the points in the overlapping region are removed from the data set when the first reg-class has been detected, to which reg-class these points eventually belong will be determined only after all reg-classes have been found. Thus, based on the rule in (34), the final result in the partitioning of reg-classes is almost independent of the extraction order.

4 Simulations

In this section, we give four numerical simulation examples and a real-life feature identification to illustrate the effectiveness and applicability of the RCMD method. Example 2 is an application of the RCMD method in the problem of switching regression models. Example 3 deals with structure mining involving the mixture of curve and line. Example 4 gives an application and generalization of the RCMD method. Example 5 involves the mining of reg-classes for a large data set with noise.

Example 2. This example considers the simple case: $m = 2, \varepsilon_1 = \varepsilon_2 = 0$. The simulation data are generated by the method given in [13]. Tests were conducted for three cases of varying parameter values. For each case, 25 samples, each of size 200, were taken according to the linear model:

TABLE 1
True Parameters for the Three Cases in Example 2

	π_1	π_2	β_{11}	β_{12}	β_{21}	β_{22}	σ_1	σ_2
Case 1	0.50	0.50	0.0	0.0	1.0	0.0	0.25	0.25
Case 2	0.50	0.50	0.0	0.0	1.0	0.0	0.75	0.75
Case 3	0.75	0.25	0.0	0.0	1.0	0.0	0.25	0.125

$$G_i : Y = \beta_{i1}X + \beta_{i2} + e_i,\ i = 1, 2, \tag{35}$$

where $X \sim U(-3, 3), e_i \sim N(0, \sigma_i^2)$, and all model parameters are given in Table 1. Each datum (X_k, Y_k) in the model G_i was generated by the following scheme:

First, an uniform random number $Z \sim U(0, 1)$ is generated, and its value is used to select a particular linear model from (35). If $Z < \pi_1$, then model G_1 is selected; otherwise, model G_2 is selected. Next, X_k is chosen to be a uniform random number in $U(0, 1)$ and a normal random variable e_i with mean 0 and standard deviation σ_i is calculated. The value Y_k is assigned using (35).

Typical scatterplots for each of the three cases are shown in Figs. 3a, 3b, and 3c depicting both the data and the optimal linear fit to each class. The GA algorithm was employed in this example with the following parameters: $P_{\text{size}} = 200$, $P_c = 0.8$, $P_m = 0.5$.

The main results of the simulation are summarized in Table 2. As a reference, we also tabulate relevant results obtained by FCRM and EM [13].

Although the scheme we used to generate the data set is the same as that in [13], the resulting difference among the RCMD method, FCRM, and EM should be regular because of the randomness of samples. Even though the results obtained by the three methods were not compared directly on the same data set, we still can observe a fact from Table 2 due to the same scheme of data generation. The fact is that for the three types of data sets in Example 2, the estimation result obtained by the RCMD method is at least not inferior to that of FCRM and EM on the average. Furthermore, our method can provide the estimation of error variances simultaneously.

To further illustrate the mining process of the RCMD method, we select randomly a sample from Case 1 for a sequential display of the procedure at work. The original sample scatterplot is shown in Fig. 4a. At the level $t = 0.1$, we can find a set of optimal preliminary estimates with $l_i(\theta_i; t_i)$, $\hat{\beta}_{11} = 0.004$, $\hat{\beta}_{12} = 0.001$, $\hat{\sigma}_1 = 0.258$. Thus, the reg-class G_1 in (26) is given by $\{(x, y) : |y - \hat{\beta}_{11}x - \hat{\beta}_{12}| \leq 3\hat{\sigma}_1\}$, which is subsequently removed from the original data set (see Fig. 4b). For the remainder data set, we again apply the RCMD method to search for another reg-class and find its parameter estimates at the same level: $\hat{\beta}_{21} = 1.003$, $\hat{\beta}_{22} = 0.016$, $\hat{\sigma}_2 = 0.225$, (see Fig. 4c). Finally, with the two classes of estimates, the data can be reclassified into two reg-classes shown in Fig. 4d, where the overlapping of the two reg-classes is classified with (34).

Generally speaking, when there are multiple reg-classes, the corresponding peak values of the objective function are different. Thus, the maximum peak value is found first. However, if the local maximum is first found or each peak value is approximately the same, then the order of extraction may be of concern. We claim that the final result is almost independent of the extraction order, meaning that the difference in the final results is very small for different extraction orders. For example, in Case 1 of Example 2, the two corresponding peaks are quite similar. In fact, at level $t = 0.1$ their mean peaks over 25 samples are -194.891 and -201.768, respectively. The rows in Table 3 represent the estimated results of different extraction order. It can be observed that the difference is indeed quite small.

In addition, we also investigated the sensitivity to initialization of the iterative algorithm. A comparison between the iterative algorithm and GA algorithm is given for Case 1 of Example 2 (see Table 4), where the initial values of the iterative algorithm selected for the reg-classes are $(\beta_{11}, \beta_{12}, \sigma_1) = (-1, 3, 4)$ and $(\beta_{21}, \beta_{22}, \sigma_2) = (2, 3, 2)$. We can observe that for good initial values these two algorithms have approximately the same results. However, in other cases, particularly those with bad initial values, the GA algorithm is preferable. It should be noted that a comparison of them is not helpful in the general sense (i.e., without considering the choice of initial values).

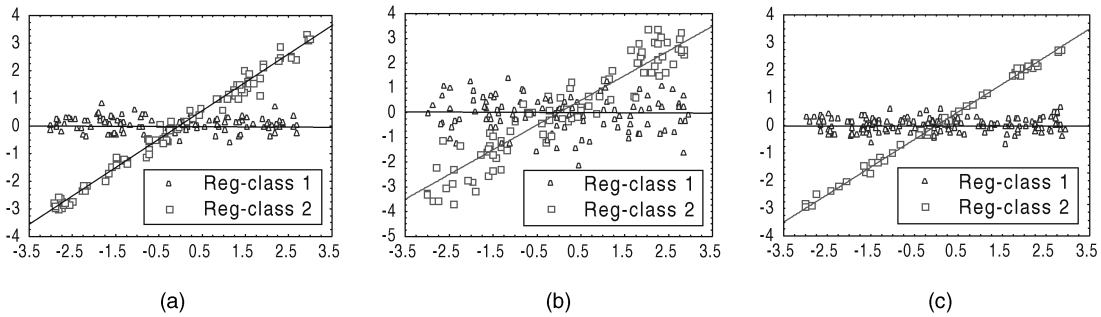

Fig. 3. Scatterplots and linear reg-classes in data sets of Example 2. (a) Case 1. (b) Case 2. (c) Case 3.

TABLE 2
Simulation Results and Comparison (on Averages Using 25 Samples of Size 200 in Each Case) for Example 2

	Case 1			Case 2			Case 3		
	Mean	Median	S. Dev.	Mean	Median	S. Dev.	Mean	Median	S. Dev.
β_{11}	0.00564	0.008	0.0196	0.00300	-0.009	0.0740	0.00384	0.005	0.0157
(FCRM)	0.00210			-0.03900			-0.00710		
(EM)	0.00480			0.01700			-0.00130		
β_{12}	-0.01092	-0.016	0.0343	0.02200	0.019	0.1317	0.00072	0.000	0.0228
(FCRM)	0.00440			0.00140			0.00170		
(EM)	0.00620			0.01300			0.00018		
σ_1	0.24832	0.256	0.0273	0.51430	0.512	0.0570	0.24956	0.256	0.0138
β_{21}	0.99596	0.995	0.0181	0.98550	0.989	0.1087	1.00360	1.003	0.0114
(FCRM)	1.00200			1.06100			0.99300		
(EM)	0.99800			1.01400			0.99800		
β_{22}	-0.00420	-0.005	0.0298	0.00050	-0.016	0.1241	0.00280	0.002	0.0237
(FCRM)	-0.00930			-0.00180			-0.00100		
(EM)	-0.00910			-0.00370			0.00039		
σ_2	0.25200	0.256	0.0310	0.55990	0.552	0.0651	0.11068	0.113	0.0175
Label.%.	7.42 7 (FCRM) 7.3 (EM)			20.07 19.8 (FCRM) 20.5 (EM)			4.98 5.9 (FCRM) 4 (EM)		

Key: Label.%.: Average label error (misclassification) in percentage using rule (34).
S.Dev.: Standard deviation.

To show the effect of a partial model t on the RCMD estimator, for a sample of Case 3 we set t from 0 to 1 by the incremental step of 0.01. One hundred preliminary parametric estimates are generated by the RCMD method. The three scatterplots depicted in Fig. 5 show that when $t = 0$, all of the parametric estimates are very poor. For $t > 0$, however, the estimates obtained by the proposed method all approximately stabilize in the vicinity of the true parameters. However, the points in Fig. 5b are not so stable. The reason may be that the degree of accuracy in estimation is not high enough. If the accuracy of optimization is increased, then the stability of the result will be increased. Another reason may result from the fact that the effect of β_2 on the absolute residual $|y - \beta_1 x - \beta_2|$ is small relatively to that of β_1. It results in that the estimation for β_2 is not as good as β_1. Even so, if we select any t in the above range for the data set, our method almost always gives a good estimate.

It should be observed that the RCMD method can be used to detect effectively breakpoints which are dominant points on the boundary of an object. This problem is important in 2D-feature extraction for computer vision. A statistical approach for detecting breakpoint has been developed based on LS method and the covariance propagation technique [17]. Clearly, the problem can be readily solved by detecting two straight lines using the RCMD method. Such a detection is robust and relatively simple.

Example 3. To demonstrate the effectiveness of the RCMD methods, the mining of linear and nonlinear structures in a mixture data set was performed in this experiment. We consider the case in which there are two reg-classes in the data set. In the simulation run, 500 data points are generated according to the following models:

$$\begin{cases} \text{Reg-class 1:} Y;=\sin(2X)-4\cos(X)+e_1, e_1\sim N(0,0.25^2), & \pi_1(1-\varepsilon_1)=0.36, \\ \text{Reg-class 2:} Y=0.4X+1+e_2\ e_2\sim N(0,0.2^2), & \pi_2(1-\varepsilon_2)=0.24, \\ \text{Noise:} X\sim U(-5,5), Y\sim U(-5,5), & \pi_1\varepsilon_1+\pi_2\varepsilon_2=0.4. \end{cases}$$

That is, there are approximately 180 (500×0.36) data points (inliers) from reg-class G_1, 120 (500×0.24) data points (inliers) from reg-class G_2, and the rest are noise. The scatterplot of the data set is depicted in Fig. 6. First, we can detect a straight line: $y = \beta_1 x + \beta_2$, at $t = 0.2$, we obtain the preliminary parametric estimation of the reg-class $G_2 : \hat{\beta}_1 = 0.395$, $\hat{\beta}_2 = 1.052$, $\hat{\sigma} = 0.199$, $l_{\max} = -483.386$. With the estimated reg-class (26) corresponding to the detected line, 152 data points are then removed from the data set. Now, we select a set of nonlinear basis functions $\{1, \sin(x), \cos(x), \sin(2x), \cos(2x)\}$ provided it is known a priori. Thus, we need to estimate the coefficients β_j in the expression:

$$Y = \beta_1 \sin(2X) + \beta_2 \cos(2X) + \beta_3 \sin X + \beta_4 \cos X + \beta_5 + e, \quad e \sim N(0, \sigma^2).$$

At $t = 0.2$, we obtain the preliminary and final estimation results summarized in Table 5. Performing the test of normality, we obtain the K-S statistic $D = 0.04065 (p > 0.20)$ and the S-W statistic $W = 0.97824 (p = 0.238471)$. The normality is thus accepted at 0.01 level of significance and the corresponding 189 data points are found.

In this example, the proportion of outliers with respect to reg-class G_2 is much more than that of the inliers. The latter is 0.24 while the former is 0.76. It is apparent that the RCMD method is very robust to the existence of outliers.

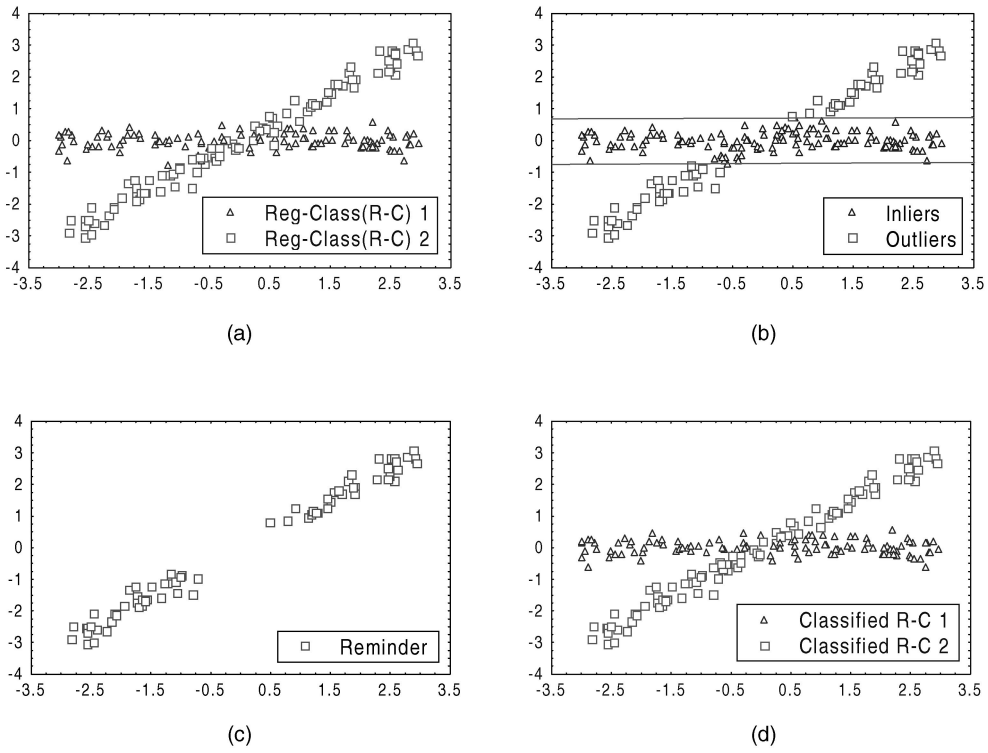

Fig. 4. Sequential mining of reg-classes by the RCMD method. (a) Original sample scatterplot. (b) Stripe associated with reg-class 1. (c) Reg-class 2 after reg-class 1 is removed. (d) Reclassified results with reg-classes.

Thus, the RCMD method appears to be effective in the learning of a variety of representations targeted by methods such as the support vector machine (SVM) [4], [31], and the dictionary methods [10].

Example 4. Curve detection is an important problem of computer vision and pattern recognition. Unraveling of quadratic curves are usually encountered in this problem. There are many approaches to detect quadratic curves. We show, in this example, that the RCMD method can be used to accomplish effectively such a task. As an illustration, we use our method to detect an ellipse with the large and small axes parallel to the coordinate axes under noise contamination.

The sample data set $\{(x_j, y_j) : j = 1, \ldots, n\}$ consists of 140 data points coming from the population (X, Y), where

$$X = x + e_1, Y = y + e_2, \frac{(x-5)^2}{3^2} + \frac{(y-4)^2}{2^2} = 1,$$
$$e_1 \sim N(0, 0.2^2),\ e_2 \sim N(0, 0.1^2),$$

and 160 data points being random noise uniformly distributed on $[0, 10] \times [0, 8]$ (Fig. 7a). In this case, the number of outliers is more than that of inliers. We select

$$r_j^2 \equiv r^2(x_j, y_j; \beta_1, \beta_2, \gamma_1, \gamma_2) \equiv 1 - \left(\frac{(x_j - \beta_1)^2}{\gamma_1^2} + \frac{(y_j - \beta_2)^2}{\gamma_2^2} \right)$$

as an error measure. The objective function can be given by

$$l(\theta) \equiv l(\beta_1, \beta_2, \gamma_1, \gamma_2; \sigma, t) \equiv \sum_{j=1}^{n} \ln \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{r_j^2}{2\sigma^2} \right) + t \right], \theta = (\beta_1, \beta_2, \gamma_1, \gamma_2, \sigma).$$

Although r_j is not necessarily normal, the function in (24) can still be used, just like the LS method which may be

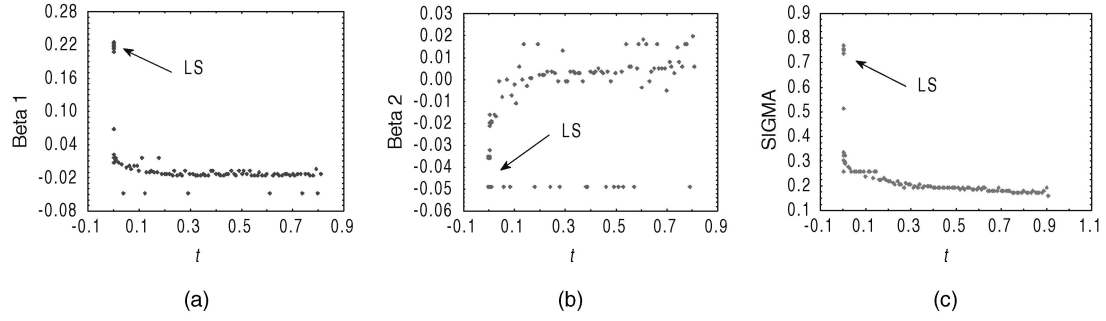

Fig. 5. Effect of partial model t on the RCMD estimator in Example 2 (Case 3). (a) Effect of t on β_1. (b) Effect of t on β_2. (c) Effect of t on σ.

TABLE 3
A Comparison for Extraction Order

	β_{11}	β_{12}	σ_1	β_{21}	β_{22}	σ_2
Pre.(1)	0.035	0.038	0.258	0.995	-0.043	0.254
Pre.(2)	0.015	0.036	0.215	0.975	-0.016	0.297
Final.(1)	0.0374	0.0413	0.2765	0.9787	-0.0135	0.3079
Final.(2)	0.0341	0.0315	0.2661	0.9669	-0.0021	0.3368

The number in parenthesis represents which reg-class is detected first.

TABLE 4
A Comparison between the Iterative and the GA Algorithms

Parameter	Mean	Median	Min	Max	Std. Dev.
β_{11}	0.00564	0.008	-0.036	0.034	0.01967
	0.01910	0.020	-0.005	0.051	0.01610
β_{12}	-0.01092	-0.016	-0.074	0.055	0.03434
	0.00101	-0.000	-0.038	0.055	0.03302
σ_1	0.24832	0.256	0.200	0.313	0.02732
	0.26515	0.259	0.225	0.356	0.03932
β_{21}	0.99596	0.995	0.963	1.028	0.01813
	0.98578	0.979	0.963	1.020	0.01642
β_{22}	-0.00420	-0.005	-0.049	0.049	0.02986
	0.00234	0.000	-0.037	0.040	0.02518
σ_2	0.25200	0.256	0.176	0.304	0.03104
	0.27553	0.279	0.244	0.304	0.02321

For each parameter, the GA results are given in the first row and the iterative algorithm results are given in the second row.

derived by the ML method. However, using the function in (24) as a parameter estimation criterion, we need to study further its theoretic meaning and interpretability.

At $t = 0.1$, we find the optimal solution:

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}) = (5.007, 3.997, 3.056, 2.010, 0.200)$$

and $l_{\max} = -366.780$. The corresponding detected ellipse is shown in Fig. 7a. Since r_j is not necessarily normal, reg-classes based on the normal error assumption are not defined. However, the t-level set given by (25) can be viewed as an inlier set supporting the detected ellipse at a partial model level t,

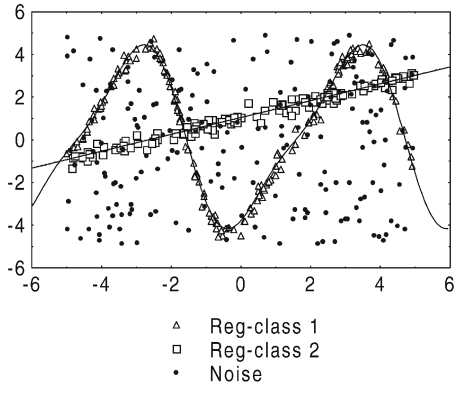

Fig. 6. Mining of linear and nonlinear structures on a mixture data set.

$$G(\hat{\theta}; 0.1) = \left\{ (x_j, y_j) : \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{\hat{r}_j^2}{2\hat{\sigma}^2}\right) > t \right\} = \left\{ (x_j, y_j) : \hat{r}_j^2 < 2\hat{\sigma}^2 \ln\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}t}\right) \right\}$$
$$= \left\{ (x_j, y_j) : 1 - 0.08392239 < \frac{(x_j - \hat{\beta}_1)^2}{\hat{\gamma}_1^2} + \frac{(y_j - \hat{\beta}_2)^2}{\hat{\gamma}_2^2} < 1 + 0.08392239 \right\},$$

which forms a stripe region (see Fig. 7b). The true ellipse is included in this stripe.

This example also shows that the RCMD method can be used to mine nonlinear patterns in large data sets with noise. Hence, the RCMD method has a great potential in its applications.

Example 5. We simulate the use of the RCMD method to mine reg-classes in a large data set with noise in this experiment. The true model is

$$G_i: \quad Z = \beta_{i1} X + \beta_{i2} Y + \beta_{i3} + e_i \,, i = 1, 2,$$

where $X \sim U(-500, 500)$, $Y \sim U(-1,000, 1,000)$, $e_i \sim N(0, \sigma_i^2)$ and all model parameters are given in Table 6. The data are generated as follows: First, random numbers $W \sim U(0, 1)$, $X \sim U(-500, 500)$, $Y \sim U(-1,000, 1,000)$, and $e_i \sim N(0, \sigma_i^2)$ are generated. Second, according to the value of W, the value Z is assigned using

$$\begin{cases} \text{Reg-class 1} \begin{cases} \text{Noise 1:} Z = X - Y + Z_1, & \text{if } W < 0.12 \\ \text{Class 1:} Z = \beta_{11} X + \beta_{12} Y + \beta_{13} + e_1, & \text{if } 0.12 \le W < 0.4 \end{cases} \\ \text{Reg-class 2} \begin{cases} \text{Noise 2:} Z = Z_2, & \text{if } 0.4 \le W < 0.64 \\ \text{Class 2:} Z = \beta_{21} X + \beta_{22} Y + \beta_{23} + e_2, & \text{if } W \ge 0.64, \end{cases} \end{cases}$$

TABLE 5
Parameteric Estimation for Example 3

Reg-classes		β_1	β_2	β_3	β_4	β_5	σ	D	W
1	Pre.	0.981	0.081	-0.023	-3.855	0.208	0.257	0.0581	0.9596
	Final LS	0.9612	0.0627	-0.0256	-3.9287	0.08587	0.2045		
	Std.Dev.	0.0226	0.0244	0.0224	0.02780	0.01824			
2	Pre.	0.395	1.052				0.199	0.0438	0.9716
	Final LS	0.3951	1.0498				0.179		
	Std.Dev.	0.0049	0.0147						

where $Z_1 \sim N(-20, 50^2)$ and $Z_2 \sim U(-1,500, 1,500)$ are random noise contained in reg-class 1 and reg-class 2, respectively. In this example, the total number of data points is $N = 10,000$. According to the scheme of assignment, $\pi_1 = 0.4, \pi_2 = 0.6, \epsilon_1 = 0.3, \epsilon_2 = 0.4$. So, theoretically the number of contaminated reg-class 1 and reg-class 2 are $\pi_1 N = 4,000$, and $\pi_2 N = 6,000$, respectively. Indeed, in this data set there are 4,056 data points in contaminated reg-class 1, where 2,830 points are inliers to reg-class 1 and 1,226 data points are Noise 1. On the other hand, there are 5,944 data points in contaminated reg-class 2, where 3,510 data points are inliers and 2,434 data points are Noise 2. We see that for each class of inliers, the number of the corresponding outliers exceeds the number of inliers.

The process for mining the reg-classes in the data set is as follows: Randomly draw from this set 10 samples with size 100. For each sample, the RCMD estimator and the parameters of the two reg-classes are obtained (Table 6). With these results, the reg-classes in the data set are respectively found and the normality test for residuals is made. The D statistic (K-S) and the number of the reg-class, n, extracted are given in columns 7 and 8 of Table 6. At significance level $\alpha = 0.01$, normality is accepted.

It is amazing that in almost all of samples, two reg-classes can be found. Although the results possess a certain degree of randomness, the effectiveness of the RCMD method for mining reg-classes in a large data set is rather convincing.

In general, when a model can be written as the form $Y = \sum_{j=1}^{p} g_j(X)\beta_j + e$, where $g_j(X)$ are fixed basis functions, it can be treated as a linear reg-class and identified by the proposed method.

Example 6. To demonstrate the practicality of the proposed algorithm, we give a real-life application here. The task is to mine line objects in remotely sensed data. In this application, a remotely sensed image from LANDSAT Thematic Mapper (TM) data acquired over a suburb area in Hangzhou, China is used to identify runways. The region contains the runway and parking apron of a certain civilian aerodrome. The image (see the left of Fig. 8) consists of a finite rectangular 95×60 lattice of pixels. To identify the runway, we use Band 5 as a feature variable. First, a feature subset of data, depicted in the right of Fig. 8, is extracted by using a simple technique which selects a pixel point when its gray-level value is above a given threshold (e.g., 250). For the lattice coordinates of points in the subset, we then perform the RCMD method to identify two runways, which can be viewed as two reg-classes. At $t = 0.05$ level, two line equations identified by the RCMD method are $y = 0.774x + 34.874$ and y = $0.341x + 22.717$, respectively. The result shows an almost complete accordance with data points in the right of Fig. 8. In other words, line-type objects such as runways and highways in remotely sensed images can easily and accurately be detected. Compared with existing techniques, such as the window method, the RCMD method can avoid the problem of selecting the appropriate window sizes and yet obtains the same results.

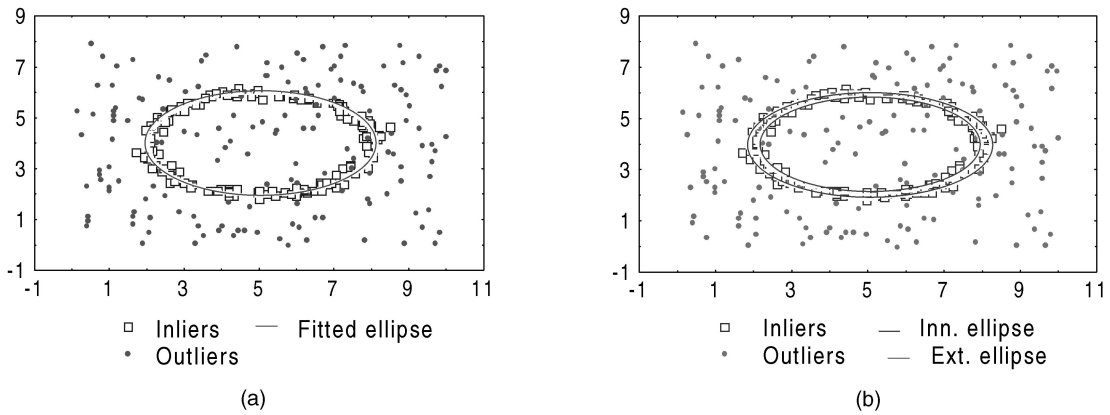

Fig. 7. Detecting an ellipse by the RCMD method. (a) Estimation result obtained by the RCMD method. (b) A stripe support set for the detected ellipse at $t = 0.1$.

TABLE 6
Parameter Estimation of 10 Samples of Size 100 in Example 5

Case	$\beta_{11}=1$	$\beta_{12}=-1$	$\beta_{13}=0$	$\sigma_1=5$	l_1	D	n
	$\beta_{21}=0$	$\beta_{22}=1$	$\beta_{23}=0$	$\sigma_2=8$	l_2		
1	1.008	-0.997	0.167	4.997	-399.438	0.03027	2827
	-0.001	1.003	-2.329	8.192	-260.865	0.03559	3282
2	1.002	-0.996	0.308	4.096	-396.560	0.03656	2698
	-0.001	1.002	-0.304	6.400	-270.051	0.03068	3085
3	1.000	-1.000	1.744	4.096	-392.260	0.04294	2861
	0.000	1.002	-0.677	6.848	-263.867	0.02739	3151
4	0.998	-1.001	2.768	5.247	-254.334	0.03873	2859
	0.000	0.993	-2.034	4.771	-395.938	0.05121	2561
5	0.999	-1.000	2.767	5.120	-253.720	0.03935	2849
	0.006	1.008	2.768	6.144	-403.747	0.06021	2789
6	1.000	-1.001	0.720	2.874	-396.720	0.04501	2430
	0.004	0.997	-1.329	6.607	-286.106	0.03782	3099
7	0.991	-1.005	0.186	3.075	-266.128	0.04818	2194
	0.004	1.001	-0.252	5.247	-407.540	0.03752	2876
8	0.998	-1.001	-0.323	5.012	-402.532	0.02983	2910
	0.016	1.000	-0.520	7.168	-264.762	0.03532	3043
9	0.992	-1.003	2.768	4.764	-405.316	0.05179	2653
	-0.016	1.000	2.768	5.824	-260.788	0.04716	2720
10	0.991	-1.002	-1.057	4.096	-396.147	0.04251	2613
	-0.002	1.002	1.871	4.352	-249.026	0.05081	2624
Mean	0.9979	-1.000	1.0048	4.3367			2689.4
	0.0010	1.0008	-0.004	6.1553			2923.0

5 Conclusion

We have proposed in this paper a new method for mining reg-classes within a general framework. The RCMD estimator has been derived and numerically verified. The robustness and effectiveness of the RCMD method have been demonstrated in a set of simulation experiments. It has been shown that the proposed method is suitable for the mining of reg-classes in small, medium, and large data sets. It appears that it is a promising method for a large variety of applications. As an effective means for data mining, the RCMD method has the following advantages:

1. The number of reg-classes does not need to be specified a priori.
2. The proportion of noisy data in the mixture can be large. Neither the number of outliers nor their distributions is part of the input (such as GMDD [33]). In this sense, the method is very robust.
3. The computation is quite fast and effective, and can be implemented by parallel computing.
4. Mining is not limited to straight lines and planes. It can also extract many curves which can be linearized (such as polynomials) and can deal with some high dimensional problems.
5. It estimates simultaneously the regression and scale parameters. Thus, the effect of the scale parameters on the regression parameters is considered. This may be more effective.

Though the RCMD method appears to be rather successful, at least by the simulation experiments, in the mining of reg-classes, there are problems, e.g., the singularity issue [30], which should be further investigated. However, with good starting values, singularities are less likely to happen. The study in [3] indicates that the incidence of singularity decreases with the increase in sample size and the increase in the angle of separation of two linear reg-classes. Obviously, we need to further study this aspect within the RCMD framework, though many researchers think that the issue of singularity in MLE may have been overblown (e.g., see [3]).

The second issue that deserves further study is the problem of sample size in the RCMD method. If a small fraction of reg-classes contains rare, but important, response variables, complications may arise. In this situation, retrospective sampling may be attempted [24]. In general, how to select a suitable sample size in RCMD is a problem which needs theoretical and experimental investigations.

To substantiate the analytical framework, theoretical properties for the RCMD estimator such as the breakdown point and exact fit point should be derived. All of the above issues constitute interesting problems for further research.

Appendix

Proof of Theorem 1. According to the strong law of large numbers, we have

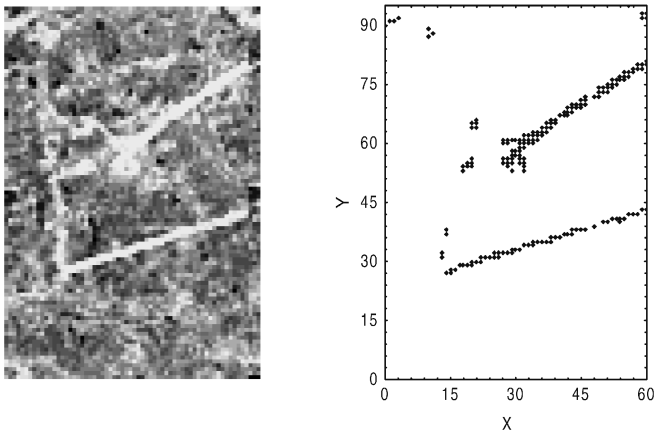

Fig. 8. Indentification of line objects in remotely sensed data.

$$\frac{1}{n}l(\theta) \xrightarrow{a.s.} E_\varepsilon[\ln p_0(X,Y;\theta)] = \iint_{R^{p+1}} \ln p_0(x,y;\theta)p_\varepsilon(x,y;\theta_0)dxdy = -I_\varepsilon(\theta;\theta_0).$$

By simple computation, we find

$$I_\varepsilon(\theta;\theta_0) = I_0(\theta;\theta_0) - \sum_{i=1}^{m}\varepsilon_i\pi_i \mathrm{B}_i(\theta).$$

Let

$$\Gamma = \{\theta \in \Theta : \theta = \theta_\varepsilon \pm k^{-1}, \text{for positive integers } k\}.$$

Then, for each $\theta \in \Gamma$, there is a set N_θ such that $P_{\theta_\varepsilon}(\bar{N}_\theta) = 0$, and when $(X,Y) \in N_\theta$,

$$\frac{1}{n}l(\theta) \to E_\varepsilon[\ln p_0(X,Y;\theta)] = -I_\varepsilon(\theta;\theta_0). \tag{A1}$$

Since Γ is countable, we have $P_\theta(N) = 0$, where $N = \bigcup_{\theta\in\Gamma}\bar{N}_\theta$. Then, when $(X,Y) \notin N$, the expression in (A1) holds for any $\theta \in \Gamma$. For any $\delta > 0$, let $\theta_1 = \theta_\varepsilon - m^{-1}$ with $m > \delta^{-1}, \theta_2 = \theta_\varepsilon + k^{-1}$ with $k > \delta^{-1}$. Then, $\theta_\varepsilon - \delta < \theta_1 < \theta_\varepsilon < \theta_2 < \theta_\varepsilon + \delta$ and for $(X,Y) \notin N$, but n large enough,

$$I_\varepsilon(\theta_1;\theta_0) > I_\varepsilon(\theta_\varepsilon;\theta_0) < I_\varepsilon(\theta_2;\theta_0) \Rightarrow l(\theta_1) < l(\theta_\varepsilon) > l(\theta_2).$$

Since $l(\theta)$ is continuous by the hypothesis, this implies that there exists a $\hat{\theta}$ in $(\theta_1,\theta_2) \subset (\theta_\varepsilon - \delta, \theta_\varepsilon + \delta)$ which maximizes $l(\theta)$. The arbitrariness of δ implies $\hat{\theta} \xrightarrow{a.s.} \theta_\varepsilon$.

The asymptotic expansion in (17) is obtained by using the Taylor formula to the left hand side of the equation $\nabla_\theta I_\varepsilon(\theta;\theta_0)|_{\theta=\theta_\varepsilon} = \mathbf{0}$. By the use of (12)-(15), we can get

$$\begin{aligned}
&\nabla_\theta I_\varepsilon(\theta;\theta_0)\big|_{\theta=\theta_\varepsilon} \\
&= \nabla_\theta I_\varepsilon(\theta;\theta_0)\big|_{\theta=\theta_0} + \nabla_\theta^2 I_\varepsilon(\theta;\theta_0)\big|_{\theta=\theta_0}(\theta_\varepsilon - \theta_0) \\
&\quad + O\Big(||\theta_\varepsilon - \theta_0||^2\Big)\mathbf{1} \\
&= \Big[\nabla_\theta I_0(\theta;\theta_0) + \sum_{i=1}^{m}\varepsilon_i\pi_i\nabla_\theta B_i(\theta)\Big]\big|_{\theta=\theta_0} \\
&\quad + \nabla_\theta^2 I_\varepsilon(\theta;\theta_0)\big|_{\theta=\theta_0}(\theta_\varepsilon - \theta_0) + O\Big(||\theta_\varepsilon - \theta_0||^2\Big)\mathbf{1} \\
&= J_\varepsilon(\theta_0)(\theta_\varepsilon - \theta_0) - \sum_{i=1}^{m}\varepsilon_i\pi_i\nabla_\theta B_i(\theta_0) + O\Big(||\theta_\varepsilon - \theta_0||^2\Big)\mathbf{1} = \mathbf{O}.
\end{aligned}$$

This completes the proof of the theorem. □

Proof of Corollary 1. From (11), we have

$$\begin{aligned}
p_\varepsilon(x,y;\theta_0) &= \sum_{i=1}^{m}\pi_i p_i^\varepsilon(x,y;\theta_i^0) \\
&= \sum_{i=1}^{m}\pi_i(1-\varepsilon_i)p_i(x,y;\theta_i^0) + \sum_{i=1}^{m}\pi_i\varepsilon_i h_i(x,y).
\end{aligned} \tag{A2}$$

If we let $\varepsilon_1 = \ldots = \varepsilon_m = \varepsilon$ and $h_i(x,y) = \Delta_{(u,v)}(x,y)$, where $\Delta_{(u,v)}(x,y)$ denotes the probability measure which puts mass 1 at the point (u,v), then the expression in (A2) becomes

$$p_\varepsilon(x,y;\theta_0) = (1-\varepsilon)p_0(x,y;\theta_0) + \varepsilon\Delta_{(u,v)}(x,y),$$

whose distribution is denoted by F_ε. Let $T(\cdot)$ denote the functional corresponding to $\hat{\theta}$. By the definition of IF and Theorem 1, it follows that

$$\begin{aligned}
&IF(u,v;\hat{\theta}) \\
&= \lim_{\varepsilon\to 0}\frac{T(F_\varepsilon) - T(F_0)}{\varepsilon} = \lim_{\varepsilon\to 0}\frac{\theta_\varepsilon - \theta_0}{\varepsilon} \\
&= \sum_{i=1}^{m}\pi_i\big[J_0(\theta_0)\big]^{-1}\nabla_\theta B_i(\theta)\Big|_{\theta=\theta_0} \\
&= \big[J_0(\theta_0)\big]^{-1}\sum_{i=1}^{m}\pi_i\nabla_\theta\Big[\iint\big(\Delta_{(u,v)}(x,y) \\
&\quad - p_i(x,y;\theta_i^0)\big)\ln p_0(x,y;\theta)dxdy\Big]\Big|_{\theta=\theta_0} \\
&= \big[J_0(\theta_0)\big]^{-1}\Big[\nabla_\theta \ln p_0(u,v;\theta) \\
&\quad - \nabla_\theta\iint p_0(x,y;\theta_0)\ln p_0(x,y;\theta)dxdy\Big]\Big|_{\theta=\theta_0} \\
&= \Big[J_0(\theta_0)\Big]^{-1}\nabla_\theta \ln p_0(u,v;\theta)\Big|_{\theta=\theta_0}.
\end{aligned}$$

This completes the proof of the corollary. □

ACKNOWLEDGMENTS

This project was supported by the earmarked grant CUHK 321/95H of the Hong Kong Research Grants Council. The authors would like to thank the referees for their valuable comments and suggestions and Dr. J.C. Luo for his assistance in the remote sensing experiment.

REFERENCES

[1] *Clustering and Classification,* P. Arabie, L.J. Hubert, and G. De Soete, eds. Singapore: World Scientific, 1996.
[2] *Intelligent Data Analysis: An Introduction,* M. Berthold and D.J. Hand, eds. New York: Springer, 1999.
[3] S.B. Caudill and R.N. Acharya, "Maximum-Likelihood Estimation of a Mixture of Normal Regressions: Starting Values and Singularities," *Comm. Statistics–Simulation,* vol. 27, no.3, pp. 667-674, 1998.
[4] V. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods.* New York: John Wiley & Sons, 1998.
[5] *Advances in Intelligent Data Analysis,* D.J. Hand, J.N. Kok, and M. Berthold, eds. Berlin: Heidelberg, Springer-Verlag, 1999.

[6] C.-N. Hsu and C.A. Knoblock, "Estimating the Robustness of Discovered Knowledge," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining,* pp. 156-161, Aug. 1995.
[7] G. Danuser and M. Stricker, "Parametric Model Fitting: From Inlier Characterization to Outlier Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 2, pp. 263-280, Feb. 1998.
[8] R.N. Dave and R. Krishnapuram, "Robust Clustering Methods: A Unified View," *IEEE Trans. Fuzzy Systems,* vol. 5, no. 2, pp. 270-293, May 1997.
[9] U. Fayyad and P. Stolorz, "Data Mining and KDD: Promise and Challenges," *Future Generation Computer Systems,* vol. 13, pp. 99-115, 1997.
[10] J.H. Friedman, "An Overview of Predictive Learning and Function Approximation," *From Statistics to Neural Networks: Theory and Pattern Recognition Applications,* V. Cherkassky, J.H. Friedman, and H. Wechsler, eds., pp. 1-61. Berlin Heidelberg: Springer-Verlag, 1994.
[11] C. Glymour, D. Madigan, D. Pregibon, and P. Symth, "Statistical Themes and Lessons for Data Mining," *Data Mining and Knowledge Discovery,* vol. 1, pp.11-28, 1997.
[12] D.J. Hand, "Data Mining: Statistics and More?" *The Am. Statistician,* vol. 52, no. 2, pp. 112-118, 1998.
[13] R.J. Hathaway and J.C. Bezdek, "Switching Regression Models and Fuzzy Clustering," *IEEE Trans. Fuzzy Systems,* vol. 1, no. 3, pp. 195-204, Aug. 1993.
[14] M. Holsheimer and M. Kersten, "A Perspective on Databases and Data Mining," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining,* pp. 150-155, 1995.
[15] J.R.M. Hosking, E.P.D. Pednault, and M. Sudan, "A Statistical Perspective on Data Mining," *Future Generation Computer Systems,* vol. 13, pp. 117-134, 1997.
[16] T. Imielinski and H. Mannila, "A Database Perspective on Knowledge Discovery," *Comm. ACM,* vol. 39, pp. 58-64, 1996.
[17] Q. Ji and R.M. Haralick, "Breakpoint Detection Using Covariance Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 845-951, Aug. 1998.
[18] G.H. John, "Robust Decision Tree: Removing Outliers from Databases," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining,* pp. 174-179, Aug. 1995.
[19] M.G. Kendall, *Kendall Advanced Theory of Statistics,* fifth ed. London: Charles Griffin, 1987.
[20] E. Kim, M. Park, S. Ji, and M. Park, "A New Approach to Fuzzy Modeling," *IEEE Trans. Fuzzy Systems,* vol. 5, no. 3, pp. 328-337, 1997.
[21] K.-N. Lau, C.-H. Yang, and G.V. Post, "Stochastic Preference Modeling within a Switching Regression Framework," *Computers Ops. Res.,* vol. 23, no. 12, pp. 1163-1169, 1996.
[22] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* New York: John Wiley & Sons, 1992.
[23] G.J. McLachlan and K.E. Basford, *Mixture Models: Inference and Applications to Clustering,* New York and Basel: Marcel Dekker Inc., 1988.
[24] R.J. O'hara Hines, "An Application of Retrospective Sampling in the Analysis of a Very Large Clustered Data Set," *J. Statistical Computer Simulation,* vol. 59, pp. 63-81, 1997.
[25] *Genetic Algorithms and Evolution Strategies in Engineering and Computer Science,* D. Quagliarella, J. P'eriaux, C. Poloni, and G. Winter, eds. England: John Wiley & Sons, 1998.
[26] R.E. Quandt and J.B. Ramsey, "Estimating Mixtures of Normal Distributions and Switching Regressions," *J. Am. Statistical Assoc.,* vol. 73, no. 364, pp. 730-738, 1978.
[27] R.A. Redner and H.F. Walker, "Mixture Densities, Maximum-Likelihood and the EM Algorithm," *SIAM Rev.,* vol. 26, no. 2, pp. 195-239, 1984.
[28] J.P. Royston, "An Extension of Shapiro and Wilk's *W* Test for Normality to Large Samples," *Applied Statistics,* vol. 31, no. 2, pp. 115-124, 1982.
[29] C.V. Stewart, "MINPRAN: A New Robust Estimator for Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 10, pp. 925-938, 1995.
[30] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley & Sons, 1987.
[31] V.N. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer-Verlag, 1995.
[32] R.D. De Veaux, "Mixtures of Linear Regressions," *Computational Statistics and Data Analysis,* vol. 8, pp. 227-245, 1989.
[33] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian Mixture Density Modeling, Decomposition, and Applications," *IEEE Trans. Image Processing,* vol. 5, no. 9, pp. 1293-1302, 1996.
[34] X. Zhuang, T. Wang, and P. Zhang, "A Highly Robust Estimator through Partial-Likelihood Function Modeling and Its Application in Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, no. 1, pp. 19-35, Jan. 1992.
[35] B. Zupan and M. Bohanec, "A Database Decomposition Approach to Data Mining and Machine Discovery," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining,* pp. 299-303, 1997.

Yee Leung received the BSSc degree in geography from The Chinese University of Hong Kong in 1972, the MA and Ph D degrees in geography and the MS degree in engineering from The University of Cororado, in 1974, 1977, and 1977, respectively. He is currently a professor of geography and chairman of the Department of Geography, research fellow of The Center for Environmental Policy and Resource Management, and deputy academic director of the Joint Laboratory for Geoinformation Science at The Chinese University of Hong Kong. He has published four monographs and more than 100 articles in international journals and book chapters. His areas of specialization cover the development and application of intelligent spatial decision support systems, spatial optimization, fuzzy sets and logic, neural networks, and evolutionary computation. Dr. Leung serves in the editorial boards of several international journals and is council member of several Chinese professional organizations.

Jiang-Hong Ma received the BS degree in mathematics from Baoji Teacher's College and the MS degree in applied mathematics from Northwestern Polytechnical University, China, in 1982 and 1988, respectively. He is currently pursuing the PhD degree in applied mathematics at Xi'an Jiaotong University and is an associate professor at Chang'an University, Xi'an. His research interests include robust statistics, pattern recognition, and data mining.

Wen-Xiu Zhang graduated in information theory, probability, and statistics from Nankai University, China, in 1967. He currently serves as dean of graduate school, vice-director of Research Centre of Applied Mathematics at Xi'an Jiaotong University, editor-in-chief of the *Journal of Engineering Mathematics* (in Chinese) and vice-editor-in-chief of *Fuzzy System and Mathematics*. His research interests include applied probability theory, set-valued stochastic process, and computer reasoning in artificial intelligence. He is the author and coauthor of more than 80 academic journal papers and 12 textbooks and research monographs.