Switching Regression Models and Fuzzy Clustering

Richard J. Hathaway and James C. Bezdek, *Fellow, IEEE*

***Abstract*— A family of objective functions, called fuzzy c-regression models, is presented which can be used to fit switching regression models to certain types of mixed data. Minimization of particular objective functions in the family yields simultaneous estimates for the parameters of c regression models, together with a fuzzy c-partitioning of the data. A general optimization approach for the family of objective functions is given and corresponding theoretical convergence results are discussed. We illustrate the new approach with two numerical examples that show how it can be used to fit mixed data to coupled linear and nonlinear models.**

***Index Terms*— EM algorithm, FCRM algorithm, fuzzy clustering, fuzzy sets, mixture distributions, regression models, switching regression.**

I. Introduction

Let $S = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ be a set of data where each independent observation $\mathbf{x}_k \in \mathbf{R}^s$ has a corresponding dependent observation $\mathbf{y}_k \in \mathbf{R}^t$. In the simplest data-fitting problem we assume that a single functional relationship between $\mathbf{x}$ and $\mathbf{y}$ holds for all the data in S. In many cases a statistical framework is imposed on this problem to account for measurement errors in the data, and a corresponding optimal solution is sought. Usually, the search for a "best" function is partially constrained by choosing the functional form of $\mathbf{f}$ in the assumed relationship

$$\mathbf{y} = \mathbf{f}(\mathbf{x}; \beta) + \varepsilon, \tag{1}$$

where $\beta \in \Omega \subset \mathbf{R}^k$ is a vector of parameters to be determined, and ε is a random vector with mean vector $\mu = \mathbf{0} \in \mathbf{R}^t$ and covariance matrix $\sum$. The definition of an optimal estimate of β depends on distributional assumptions made about ε, and the set Ω of feasible values of β. This type of model is well known and can be found in most texts on multivariate statistics.

The type of model considered here is known as a switching regression model, and is discussed in varying detail in [1]–[10]. Instead of assuming that a single model can account for all n pairs in S, we assume the data to be drawn from c models:

$$\mathbf{y} = \mathbf{f}_i(\mathbf{x}; \beta_i) + \varepsilon_i, \qquad 1 \le i \le c, \tag{2}$$

where each $\beta_i \in \Omega_i \subset \mathbf{R}^{k_i}$, and each ε_i is a random vector with mean vector $\mu_i = \mathbf{0} \in \mathbf{R}^t$ and covariance matrix $\sum_i$. Good estimates for the parameters $\{\beta_1, \cdots, \beta_c\}$ are desired as in the single model case. However, we have the added difficulty that S is *unlabeled*; that is, for a given datum $(\mathbf{x}_k, \mathbf{y}_k)$, it is not known which model from (2) applies. The purpose of this paper is to present a new approach, based on fuzzy clustering techniques, that will produce estimates of $\{\beta_1, \cdots, \beta_c\}$ and at the same time assign a fuzzy label vector to each datum in S.

Applications of switching regression models are found in economics [4], [7], [8]. We give an example from fisheries research to illustrate the basic idea. According to Hosmer [5], for a certain range of ages, the mean length of a male halibut is approximately a linear function of its age. Likewise, the mean length of female halibut is approximately a linear function of their age. The most readily obtainable data that can be used to establish these relationships come from fish that have been cleaned. Each such fish yields age and length measurements, but its sex is indistinguishable. The parameters defining the linear growth curves for the two sexes can be estimated by treating the data analysis as a switching regression problem. For this example, $c = 2$, $s = 1$, $t = 1$, y = length, x = age, and the two models are

$$y = f_1(x, \beta_1) + \varepsilon_1 = \beta_{11}x + \beta_{12} + \varepsilon_1, \tag{3a}$$

$$y = f_2(x, \beta_2) + \varepsilon_2 = \beta_{21}x + \beta_{22} + \varepsilon_2. \tag{3b}$$

One approach for estimating the parameters $\{\beta_{ij}\}$ in (3) can be formulated using mixture distributions, which are described in detail in [10]. Using the simple case of (3) as an example, each piece of data (x_k, y_k) is viewed as coming from regime 1 with probability α, and from regime 2 with probability $1-\alpha$, where regimes 1 and 2 correspond to the models in (3a) and (3b), respectively. To continue, assumptions about the distributions of ε_1 and ε_2 must be made. Commonly made assumptions are that values of ε_1 and ε_2 are independent for different data (x_k, y_k) and (x_j, y_j), and that the distributions of ε_1 and ε_2 are normal with mean 0 and (unknown) standard deviations σ_1 and σ_2, respectively. Denoting the univariate normal probability density function with mean μ and standard deviation σ by

$$p(\varepsilon; \mu, \sigma) = \frac{e^{\frac{-(\varepsilon-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}},$$

we have the following log-likelihood function of the samples in S:

$$\begin{aligned} &L(\alpha, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma_1, \sigma_2; S) \\ &\quad = \sum_{k=1}^{n} \log(\alpha p(y_k - \beta_{11}x_k - \beta_{12}; 0, \sigma_1) \\ &\quad + (1-\alpha) p(y_k - \beta_{21}x_k - \beta_{22}; 0, \sigma_2)). \end{aligned} \tag{4}$$

Manuscript received November 5, 1992; revised April 10, 1993. This work was supported in part by the Faculty Research Committee of Georgia Southern University and by the NSF under grant IRI-9003252.

R. J. Hathaway is with the Mathematics and Computer Science Department, Georgia Southern University, Statesboro, GA 30460.

J. C. Bezdek is with the Department of Computer Science, University of West Florida, Pensacola, FL 32514.

IEEE Log Number 9211235.

The EM algorithm [9], or similar statistical approaches given in [1]–[3], and [5]–[8] can be used to iteratively optimize L. The approach taken here is more akin to fuzzy cluster analysis than statistics. The main problem is that the data in S are unlabeled, so numerical methods for estimation of the parameters almost always lead to equations which are coupled across classes. If S were partitioned into c *crisp* (conventional) subsets corresponding to the regimes represented by the models in (2), then estimates for $\{\beta_1, \cdots, \beta_c\}$ could be obtained by simpler methods. In particular, an alternative to the approach represented by (4) is to first find a crisp c partition of S (find c subsets of $S = \cup S_i, S_i \cap S_j = \emptyset$ for $i \neq j$) using any crisp (conventional) clustering algorithm such as hard c-means [11] and then to solve c separate single-model problems using S_i with (1).

A third alternative is to formulate the two problems (partition S and estimate $\{\beta_1, \cdots, \beta_c\}$) so that a simultaneous solution can be attempted; this is the approach we take. Towards this end, we seek a clustering criterion that explicitly accounts for both the form of the regression models and the need to partition the unlabeled data so that each cluster of S is well fit by a single model from (2). The purpose of this paper is to describe a clustering algorithm that partitions the data and simultaneously provides estimates of the parameters $\{\beta_1, \cdots, \beta_c\}$ which define the best-fit regression models.

The objective functions which define the fuzzy c-regression models (FCRM) approach are given in Section II. A general iterative scheme for minimizing particular members of the family of objective functions is developed, and certain aspects of the algorithms' convergence theory are discussed. Section III contains the results of two numerical simulations. Example 1 in that section compares fits to $c = 2$ mixed linear models made by FCRM with maximum likelihood estimates of the mixture density parameters using the EM method described above. Example 2 illustrates the behavior of FCRM when estimating the parameters of a pair of mixed quadratic models. Section IV contains our conclusions and poses some questions for future research.

II. Fuzzy C-RegressionModels

To characterize solution spaces for clustering, let c denote the number of clusters, $1 < c < n$, and set

$$\begin{aligned} E_{fcu} &= \{\mathbf{u} \in \mathcal{R}^c | u_i \in [0,1] \forall i\} \\ &= \text{(unconstrained labels);} \qquad (5a) \\ E_{fc} &= \{\mathbf{u} \in E_{fcu} | \sum_{i=1}^{c} u_i = 1\} \\ &= \text{(constrained labels);} \qquad (5b) \\ E_c &= \{\mathbf{u} \in E_{fc} | u_i \in \{0,1\} \forall i\} \\ &= \text{hard (nonfuzzy, or crisp) labels.} \qquad (5c) \end{aligned}$$

Fig. 1 depicts these sets for three classes. E_c is the canonical (unit vector) basis of Euclidean c space. E_{fc}, a subset of a hyperplane, is its convex hull; and E_{fcu} is the unit hypercube in $\Re^c$.

The set $S = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ is called a set of object data. The pair $(\mathbf{x}_k, \mathbf{y}_k)$ is a *feature vector* in $\mathbf{R}^s \times \mathbf{R}^t$ that represents the kth object which generated the data. In our fish example, each Halibut is the real object, and has a representation as a feature vector $(\mathbf{x}_k, \mathbf{y}_k)$ in $\mathbf{R}^1 \times \mathbf{R}^1$. *Clustering in S* means assigning label vectors from one of the sets in (5) to each feature vector in S. Implicitly, these labels are also assigned to the objects that they represent. The vector $\mathbf{u} = (0.1,\ 0.6,\ 0.3)^T$ is a constrained label vector; its entries lie between 0 and 1, and are constrained to sum to 1. If $\mathbf{u}$ is generated by, say, the fuzzy c-means clustering method, we call it a fuzzy label and interpret its values as the membership of $\mathbf{u}$ (and of the object $\mathbf{u}$ represents) in each of the classes represented by the rows of $\mathbf{u}$. Thus, 0.6 is the membership of $\mathbf{u}$ (and of the object $\mathbf{u}$ represents) in class 2. If $\mathbf{u}$ came from a method such as maximum likelihood estimation in mixture decomposition, it would be a probabilistic label, and 0.6 would be the (posterior) probability that $\mathbf{u}$ (and of the object $\mathbf{u}$ represents) came from class 2.

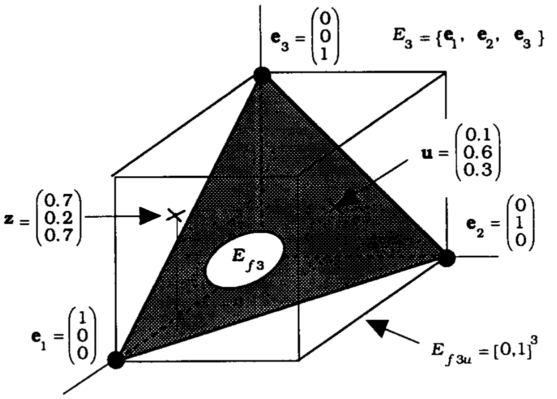

$\mathbf{e}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$
$E_3 = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$
$\mathbf{u} = \begin{pmatrix}0.1\\0.6\\0.3\end{pmatrix}$
$\mathbf{z} = \begin{pmatrix}0.7\\0.2\\0.7\end{pmatrix}$
$\mathbf{e}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$
E_{f3}
$\mathbf{e}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$
$E_{f3u} = [0,1]^3$

Fig. 1. Hard, fuzzy and probabilistic label vectors (for $c = 3$ classes).

The cube $E_{fcu} = [0,1]^3$ is called unconstrained label vector space because vectors such as $\mathbf{z} = (0.7,\ 0.2,\ 0.7)^T$ have each entry between 0 and 1, but are otherwise unrestricted. Here is a physical situation that seems best modeled by *unconstrained* label vectors. Suppose we want to assign labels (membership values) to a set of people in two fuzzy sets: scientists and artists. The membership of individuals in either fuzzy set should increase in proportion to their ability. Artistic talent does not necessarily decrease as scientific ability increases, or conversely. However, using label vectors whose entries sum to 1 would force this unnatural property on the model. In this situation the most plausible memberships are unconstrained labels. For example, daVinci might be assigned the memberships (0.6, 0.8), Michaelangelo (0.2, 0.8), and Feynman (0.9, 0.4).

Given any finite set of unlabeled data, the problem of *clustering* in S is to assign the objects (hard or fuzzy or probabilistic) labels that identify "natural subgroups" in S. This problem is sometimes called unsupervised learning: good introductions to many clustering algorithms are found in [12] and [13].

Let (c) be an integer, $1 < c < n$, and let $S = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$ denote a set of (n) *unlabeled* feature vectors in $\mathbf{R}^p$. In the case of switching regression, $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)^T$ and

TABLE I

Typical 2-Partitions of $S = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ = { Peach, Plum,Nectarine }

Object	$U_1 \in M_{23}$	$U_2 \in M_{f23}$	$U_3 \in M_{f23u}$
	$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$	$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$	$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$
Peaches	1 0 0	0.9 0.2 0.4	0.9 0.5 0.5
Plums	0 1 1	0.1 0.8 0.6	0.6 0.8 0.7

$p = s + t$. Label vectors assigned to each object in a set of data can be conveniently arrayed as $(c \times n)$ *c-partitions* of S, which are characterized as sets of (cn) values $\{U_{ik}\}$ satisfying some or all of the following conditions:

$$0 \le U_{ik} \le 1 \quad \forall i, k; \tag{6a}$$

$$0 < \sum_{k=1}^{n} U_{ik} < n \quad \forall i; \tag{6b}$$

$$\sum_{i=1}^{c} U_{ik} = 1 \quad \forall k. \tag{6c}$$

Using equations (6) with the values $\{U_{ik}\}$ arrayed as a $(c \times n)$ matrix $U = [U_{ik}]$, we define

$$M_{fcnu} = \{U \in \mathbf{R}^{cn} | U \text{ satisfies (6a) and (6b)}\}; \tag{7a}$$

$$M_{fcn} = \{U \in M_{fcnu} | U \text{ satisfies (6c)}\}; \tag{7b}$$

$$M_{cn} = \{U \in M_{fcn} | U_{ik} = 0 \text{ or } 1 \quad \forall i, k\}. \tag{7c}$$

Equations (7a), (7b), and (7c) define, respectively, the sets of unconstrained, constrained, and crisp *c-partitions* of S. Each column of U in M_{fcnu} (M_{fcn}, M_{cn}) is a label vector from E_{fcu} (E_{fc}, E_c). The reason these matrices are called partitions follows from the interpretation of U_{ik}. If U is a fuzzy c-partition, U_{ik} is taken as the *membership* of $\mathbf{z}_k$ in the ith fuzzy subset (cluster) of S. If U is a statistically derived partition, U_{ik} is usually the *posterior probability*, given $\mathbf{z}_k$, that it came from class i.

M_{fcnu} and M_{fcn} can be more realistic physical models than M_{cn}, for the boundaries between many classes of real objects (e.g., the three forms of sulfur) are badly delineated (i.e., really fuzzy). We give an example to illustrate the three kinds of partitions. Let $S = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ = {peach, plum, nectarine}, and let $c = 2$. Typical 2-partitions of these three objects are shown in Table I. The nectarine, $\mathbf{z}_3$, is shown as the last column of each partition, and in the hard case, it must be (erroneously) given full membership in one of the two crisp subsets partitioning this data; in $U_1, \mathbf{z}_3$ is labeled "plum." Noncrisp partitions enable algorithms to (sometimes!) avoid such mistakes. The final column of U_2 allocates most (0.6) of the membership of $\mathbf{z}_3$ to the plum class; but also assigns a lesser membership (0.4) to $\mathbf{z}_3$ as a peach. U_3 illustrates an unconstrained set of membership assignments for the three fruits. Columns such as the one for the nectarine in U_2 and U_3 serve a useful purpose—lack of strong membership in a single class is a signal to "take a second look." Hard partitions of data cannot suggest this. The nectarine is a *hybrid* of peaches and plums, so the labels shown for it in the last column of U_2 or U_3 seem more plausible *physically* than crisp assignment of $\mathbf{z}_3$ to either (incorrect) class. Algorithms that yield clusters in unconstrained label space M_{fcnu} are fairly new [15]. Our discussion concerns only constrained labels.

For the switching regression problem, we interpret U_{ik} as the importance or weight attached to the extent to which the model value $\mathbf{f}_i(\mathbf{x}_k; \beta_i)$ matches $\mathbf{y}_k$. Crisp memberships (0's and 1's) in this context would place all of the weight in the approximation of $\mathbf{y}_k$ by $\mathbf{f}_i(\mathbf{x}_k; \beta_i)$ on one class for each k. But fuzzy partitions enable us to represent situations where a data point fits several models equally well, or more generally, may fit all c models to varying degrees.

The final piece of notation needed to define the FCRM objective functions concerns the error in a particular model's predicted value. Specifically, let $E_{ik} : \mathbf{R}^{k_i} \mapsto [0, \infty)$ be

$$E_{ik}(\beta_i) \equiv \text{measure of error in } \mathbf{f}_i(\mathbf{x}_k; \beta_i) \text{ as an approximation to } \mathbf{y}_k, 1 \le i \le c; 1 \le k \le n. \tag{8a}$$

The most common example for such a measure is the squared vector norm $E_{ik}(\beta_i) = \|\mathbf{f}_i(\mathbf{x}_k; \beta_i) - \mathbf{y}_k\|^2$. We leave the precise nature of the measure in (8a) unspecified in order to allow a very general framework. However, all choices for E_{ik} are required to satisfy the following *minimizer property.* Let $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$ with $a_i \ge 0 \forall i$, and $\mathbf{E}_i(\beta_i) = (E_{i1}(\beta_i), \cdots, E_{\text{in}}(\beta_i))^T, 1 \le i \le c$. We require that each of the c functions (Euclidean dot products)

$$< \mathbf{a}, \mathbf{E}_i(\beta_i) > = a_1 E_{i1}(\beta_i) + a_2 E_{i2}(\beta_i) + \; + a_n E_{\text{in}}(\beta_i); 1 \le i \le c \tag{8b}$$

has a global minimum over Ω_i, the set of feasible values of β_i.

The general family of fuzzy c regression models objective functions is defined, for $U \in M_{fcn}$ and $(\beta_1, \cdots, \beta_c) \in \Omega_1 \times \Omega_2 \times \cdots \times \Omega_c \subset \mathbf{R}^{k_1} \times \mathbf{R}^{k_2} \times \cdots \times \mathbf{R}^{k_c}$, by

$$E_m(U, \{\beta_i\}) = \sum_{k=1}^{n} \sum_{i=1}^{n} U_{ik}^m E_{ik}(\beta_i), \tag{9}$$

where $m > 1$ is fixed, and the $\{E_{ik}(\beta_i)\}$ defined by (8a) satisfy the property (8b). This family is similar to the fuzzy c-varieties objective function family given in [11], but differs in that the fit of different regression models to each $\mathbf{y}_k$ replaces the distance of object data vector k to some prototype of cluster i as the measure of goodness of fit. The key assumption is that minimizers $(U^*, \beta_1^*, \cdots, \beta_c^*)$ of $E_m(U, \beta_1, \cdots, \beta_c)$ are such that U^* is a reasonable fuzzy partitioning of S and $\{\beta_1^*, \cdots, \beta_c^*\}$ determine a good switching regression model.

Clustering techniques based on minimizing functions similar to (9) are known to produce reasonable estimates of statistical parameters [16]. Our approach for minimizing (9) is to apply grouped coordinate minimization as given in [17] and [18] to E_m. We state the version where exact coordinate minimization is possible.

Fuzzy c-Regression Models (FCRM) Algorithms (Exact Inner Minimizations)

Step 1. Given data $S = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$. Set $m > 1$, specify regression models (2), and choose a measure of error $E = \{E_{ik}\}$ so that $E_{ik}(\beta_i) \geq 0$ for i and k, and for which the minimizer property (8b) holds. Pick a termination threshold $\varepsilon > 0$ and an initial partition $U^{(0)} \in M_{fcn}$. Then for $r = 0, 1, 2, \cdots$:

Step 2. Calculate values for the c model parameters $\beta_i = \beta_i^{(r)}$ that globally minimize (over $\Omega_1 \times \Omega_2 \times \times \Omega_c$) the restricted function

$$\Psi(\beta_1, \cdots, \beta_c) \equiv E_m(U^{(r)}, \beta_1, \cdots, \beta_c). \qquad (10)$$

Step 3. Update $U^{(r)} \to U^{(r+1)} \in M_{fcn}$, with $E_{ik} = E_{ik}(\beta_i^{(r)})$, to satisfy

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{E_{ik}}{E_{jk}}\right)^{\frac{1}{m-1}}} \text{ if } E_{ik} > 0 \text{ for } 1 \leq i \leq c; \qquad (11a)$$

otherwise,

$$U_{ik} = 0 \text{ if } E_{ik} > 0, \text{ and}$$
$$U_{ik} \in [0, 1] \text{ with } (U_{1k} + \cdots + U_{ck}) = 1. \quad (11b)$$

Step 4. Check for termination in some convenient matrix norm. If $\|U^{(r)} - U^{(r+1)}\| \leq \varepsilon$, then stop; otherwise set $r = r + 1$ and return to step 2.

We remark that a solution to the switching regression problem with mixture decomposition using the EM algorithm can be regarded as the same optimization approach, but applied to the objective function $L(U, \gamma_1, \cdots, \gamma_c) = \sum_{k=1}^{n} \sum_{i=1}^{c} U_{ik}(E_{ik}(\gamma_c) + \log(U_{ik}))$ (see [17, eq. (11)]). In this case, the $\{\gamma_i\}$ are the regression model parameters (the $\{\beta_i\}$), plus additional parameters such as means, covariance matrices, and mixing proportions associated with the c components of the mixture.

In our experience a choice for ε in the range 0.0001 to 0.00001 usually yields good estimates (when FCRM terminates at the right place!) in a relatively small number of iterations. In practice the implementation of step 3 can be simplified by adding a small positive number (10^{-100} was used in Example 2 in Section III) to any zero values of E_{ik} that are encountered. This does not affect the convergence behavior of the iterate sequence, and allows all membership values to be defined using (11a).

Step 2 is possible since the measure of error satisfies the minimizer property and Ψ can be rewritten to look like a sum of functions of the form in (8b). For a specific example of step 2, suppose that $t = 1$, and for $1 \leq i \leq c : k_i = s, \Omega_i = \mathbf{R}^s, \mathbf{f}_i(\mathbf{x}_k; \beta_i) = (\mathbf{x}_k)^T \beta_i$, and

$$E_{ik}(\beta_i) = (y_k - (\mathbf{x}_k)^T \beta_i)^2. \qquad (12)$$

The objective function $E_m(U, \beta_1, \cdots, \beta_c)$ for (12) is a fuzzy, multimodel extension of the least squares criterion for model fitting, and any existing software for solving weighted least squares problems can be used in step 2. The explicit formulas for the new iterates $\beta_i^{(r)}$, $1 \leq i \leq c$, can be easily derived using calculus. Let X denote the matrix in $R^{n \times s}$ having $\mathbf{x}_k$ as its kth row, Y denote the vector in $\mathbf{R}^n$ having y_k as its kth component, and D_i denote the diagonal matrix in $R^{n \times n}$ having $(U_{ik}^{(r)})^m$ as its kth diagonal element. If the columns of X are linearly independent and $U_{ik}^{(r)} > 0$ for $1 \leq k \leq n$, then

$$\beta_i^{(r)} = [X^T D_i X]^{-1} X^T D_i Y. \qquad (13)$$

If the columns of X are not linearly independent, then $\beta_i^{(r)}$ can still be calculated directly, but techniques based on orthogonal factorizations of X should be used. Though it rarely occurs in practice, $U_{ik}^{(r)}$ can equal 0 for some values of k, but this will cause singularity of $[X^T D_i X]$ only in degenerate (and extremely unusual) cases. As a practical matter, $\beta_i^{(r)}$ in (13) will be defined throughout the iteration if the columns of X are linearly independent.

The remainder of this section is concerned with theoretical convergence properties of this family of algorithms. Using LaGrange multiplier theory, it is easily shown that for $E_{ik} \geq 0$, (11) defines $U^{(r+1)}$ to be a global minimizer of the restricted function

$$\Gamma(U) \equiv E_m(U, \beta_1^{(r)}, \cdots, \beta_c^{(r)}) \qquad (14)$$

over M_{fcn}. From this it follows that iteration between (11) and (12) is a special case of grouped coordinate minimization, and the general convergence theory from [17]–[19] can be applied. Readers interested in the technical details should see these references; here, an informal statement of the main results is given. Global convergence theory from [19] can be applied for reasonable choices of $E_{ik}(\beta_i)$ in (8a) to show that any limit point of an iteration sequence will be a minimizer, or at worst a saddle point, of $E_m(U, \beta_1, \cdots, \beta_c)$. The global convergence results are derived from the monotonic descent property

$$E_m(U^{(r+1)}, \beta_1^{(r+1)}, \cdots, \beta_c^{(r+1)}) \leq E_m(U^{(r)}, \beta_1^{(r)}, \cdots, \beta_c^{(r)}). \qquad (15)$$

The local convergence result in [17] states that if the error measures $\{E_{ik}(\beta_i)\}$ are sufficiently smooth and a standard convexity property holds at a minimizer $(U^*, \beta_1^*, \cdots, \beta_c^*)$ of E_m, then any iteration sequence started with $U^{(0)}$ sufficiently close to U^* will converge to $(U^*, \beta_1^*, \cdots, \beta_c^*)$. Furthermore, the *rate* of convergence of the sequence will be q-linear. This means that there is a norm $\| * \|$, and constants $0 < \gamma < 1$ and $r_0 > 0$, such that for all $r \geq r_0$, the sequence of errors $\{e^r\} = \{\|(U^{(r)}, \beta_1^{(r)}, \cdots, \beta_c^{(r)}) - (U^*, \beta_1^*, \cdots, \beta_c^*)\|\}$ satisfies the inequality

$$e_{r+1} \leq \gamma \, e_r. \qquad (16)$$

The level of computational difficulty in executing step 2 of the algorithm is a major consideration in choosing the particular measure of error $E_{ik}(\beta_i)$. The best situation is when a closed-form solution for the new iterate $\beta_i^{(r)}$ exists such as in the example at (13). Fortunately, in cases where the minimization of $\Psi(\beta_1, \cdots, \beta_c)$ must be done iteratively, the convergence theory of [18] shows that a single step of Newton's method on Ψ, rather than exact minimization, is sufficient to preserve the local convergence results. The case of inexact minimization in each half step is further discussed and exemplified in connection with a different algorithm in [20].

III. Numerical Examples

We give two numerical examples to illustrate fuzzy c-regression models. Example 1 concerns $c = 2$ linear models, while Example 2 discusses $c = 2$ quadratic models.

Example 1

This example considers the simple $c = 2$ switching regression model given by (3). The distribution of the artificial test data and the simulation results are described after precise iterate formulas are given for both the FCRM and EM approaches. FCRM iteration is developed by adapting the general algorithm stated earlier to $E_{ik}(\beta_i) = (y_k - f_i(x_k;\beta_i))^2$ and $m = 2$, for which the calculation in step 2 becomes, for $i = 1$ and 2,

$$\beta_{i1} = \left(\left(\sum_{k=1}^{n}(U_{ik})^2\right)\left(\sum_{k=1}^{n}(U_{ik})^2 x_k y_k\right) - \left(\sum_{k=1}^{n}(U_{ik})^2 x_k\right)\left(\sum_{k=1}^{n}(U_{ik})^2 y_k\right)\right)\Big/K_i, \tag{17a}$$

and

$$\beta_{i2} = \left(\left(\sum_{k=1}^{n}(U_{ik}x_k)^2\right)\left(\sum_{k=1}^{n}(U_{ik})^2 y_k\right) - \left(\sum_{k=1}^{n}(U_{ik})^2 x_k\right)\left(\sum_{k=1}^{n}(U_{ik})^2 x_k y_k\right)\right)\Big/K_i, \tag{17b}$$

where

$$K_i = \left(\left(\sum_{k=1}^{n}(U_{ik})^2\right)\left(\sum_{k=1}^{n}(U_{ik}x_k)^2\right) - \left(\sum_{k=1}^{n}(U_{ik})^2 x_k\right)\left(\sum_{k=1}^{n}(U_{ik})^2 x_k\right)\right), \text{ i=1,2.} \tag{17c}$$

Similarly, the EM algorithm in the mixture density approach takes the form of a successive iteration between a posterior probability matrix $U \in M_{fcn}$ and parameter estimates for the model. Instead of (11) in step 3, the next matrix $U^{(r+1)}$ is calculated from the newest parameter estimate $(\alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \sigma_1, \beta_{21}, \beta_{22}, \sigma_2)$ by

$$U_{ik}^{(r+1)} = \alpha_i p(y_k - \beta_{i1}x_k - \beta_{i2}; 0, \sigma_1)/P_k \text{ for } i = 1, 2, \text{ where } \alpha_1 = \alpha, \alpha_2 = 1 - \alpha, \text{ and} \tag{18a}$$

$$P_k = \alpha_1 p(y_k - \beta_{11}x_k - \beta_{12}; 0, \sigma_1) + \alpha_2 p(y_k - b_{21}x_k - \beta_{22}; 0, \sigma_2). \tag{18b}$$

TABLE II
True Parameter Values for the Three Cases Illustrated in Example 1

	α	β_{11}	β_{12}	β_{21}	β_{22}	σ_1	σ_2
Case 1	0.50	0.0	0.0	1.0	0.0	0.25	0.250
Case 2	0.50	0.0	0.0	1.0	0.0	0.75	0.750
Case 3	0.75	0.0	0.0	1.0	0.0	0.25	0.125

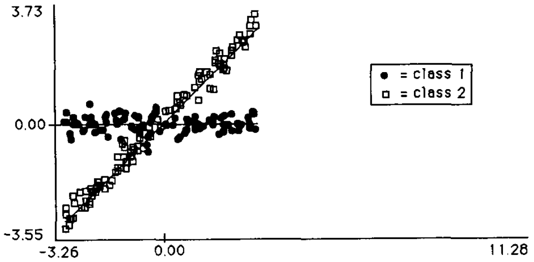


Fig. 2. Scatter plot for a case 1 sample. Graph of true model($f_1(x) = 0$; $f_2(x) = x$) included.

The equations which use the current matrix U to define the next EM parameter estimate for $\beta_{11}, \beta_{12}, \beta_{21}$, and β_{22} are obtained from (17) by replacing each occurrence of $(U_{ik})^2$ with U_{ik}. These new values of the model parameters are then used to calculate the new standard deviations:

$$\sigma_1 = \sqrt{\frac{\sum_{k=1}^{n} U_{ik}(y_k - \beta_{i1}x_k - \beta_{i2})^2}{\sum_{k=1}^{n} U_{ik}}} \quad \text{, i=1, 2.} \tag{19}$$

Finally, the new estimate of α is given by (recall that $c = 2$ here)

$$\alpha = \frac{\sum_{k=1}^{n} U_{1k}}{n}. \tag{20}$$

For purposes of comparison, (20) is also used with the terminal fuzzy c-regression models partition to calculate a fuzzy analogue of the proportion parameter. Both methods used $\varepsilon = 0.0001$ in the termination check (step 4) that stops the iteration as soon as

$$\sum_{k=1}^{n}\sum_{i=1}^{c} |U_{ik}^{(r+1)} - U_{ik}^{(r)}| < \varepsilon. \tag{21}$$

Tests were conducted for the three sets of parameter values that are given in Table II. For each of the three cases, 25 samples, each of size 200, were generated according to the model in (3); each datum (x_k, y_k) was generated by the following scheme. First, a uniform (in (0,1)) random number z_1 is generated, and its value is used to select a particular linear model from (3). If $z_1 < \alpha$, then model 1 is used; otherwise, model 2 is used. Let $i = 1$ if $z_1 < \alpha$; otherwise $i = 2$. Next, x_k is picked to be a uniform random number in (−3,3) and a normal random variate ε_i with mean 0 and standard deviation σ_i is calculated. The value y_k is assigned using (3), x_k, ε_i, and the appropriate model parameters from Table II.

Typical scatterplots for each of the three types of samples are shown in Figs. 2–4. The uniform (pseudo-) random data were generated using the Microsoft Basic random number

TABLE III
SIMULATION AVERAGES USING 25 SAMPLES OF SIZE200 FOR EACH CASE

	Case 1		Case 2		Case 3	
	FCRM	EM	FCRM	EM	FCRM	EM
# of Iterations	8.6	10.0	19.9	77.3	8.4	8.8
% Label Errors	7.0	7.3	19.8	20.5	5.9	4.0
$\hat{\alpha}$ from (20)	.502	.501	.499	.504	.685	.745
$(\alpha - \hat{\alpha})^2$	.00097	.0013	.0006	.0055	.0053	.00099
$\hat{\beta}_{11}$	.0021	.0048	−.039	.017	−.0071	−.0013
$(\beta_{11} - \hat{\beta}_{11})^2$	.00012	.00015	.0029	.0049	.00012	.000083
$\hat{\beta}_{12}$	.0044	.0062	.0014	.013	.0017	.00018
$(\beta_{12} - \hat{\beta}_{12})^2$	.00085	.00083	.013	.013	.00044	.00035
$\hat{\beta}_{21}$	1.002	0.998	1.061	1.014	.993	.998
$(\beta_{21} - \hat{\beta}_{21})^2$	.00018	.00018	.0006	.00045	.00024	.00018
$\hat{\beta}_{22}$	−.0093	−.0091	−.0018	−.0037	−.001	.00039
$(\beta_{22} - \hat{\beta}_{22})^2$	.0011	.0011	.01	.0068	.00033	.00026

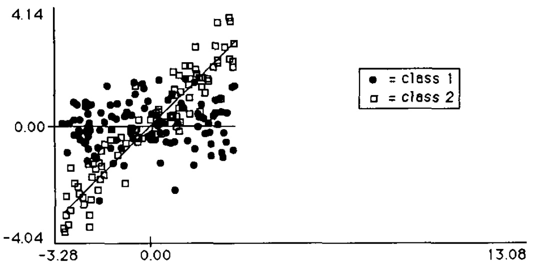


Fig. 3. Scatter plot for a Case 2 sample. Graph of true model($f_1(x) = 0$; $f_2(x) = x$) included.

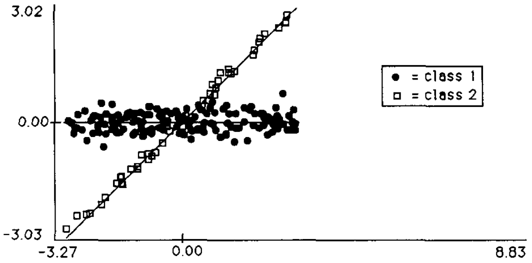


Fig. 4. Scatter plot for a Case 3 sample. Graph of true model($f_1(x) = 0$; $f_2(x) = x$)included.

TABLE IV
NUMBER OF TIMES OUT OF 25 TRIALS THAT TERMINATION OCCURRED AT POINTS OTHER THAN THE GLOBAL MINIMIZER WHEN STARTED WITH POOR INITIALIZATION

	Case 1	Case 2	Case 3
FCRM	5 (20%)	9 (36%)	3 (12%)
EM/Mixture	0	0	0

generator, and then shuffled using the technique from [21]. The normal (pseudo-) random data were then obtained using the technique proposed by Box and Muller [22].

Figs. 2–4 display both the data and the optimal (generating) linear fit to each class. Also shown with each plot are the formulas of the pair of true linear functions that would be used if the parameters $\{\beta_{ij}\}$ were known. From Table II we see that the "true but unknown" models are $f_1(x) = 0+0\cdot x \equiv 0$; $f_2(x) = 0 + 1 \cdot x \equiv x$. We compare the mixture and fuzzy c-regression models estimates of the coefficients of f_1 and f_2 below. In the computational examples, of course, the class labels of each data point are not known to the algorithms, which "see" all the data simply as points $(x_k, y_k) \in \mathbf{R}^2$.

The main results of the simulation are summarized in Table III. In this table, fuzzy c-regression models and EM were both initialized using the correct hard partitions of the data, which were recorded during data generation and are represented in the examples of Fig. 2–5 by the symbols • = class 1; □ = class 2. Iteration counts in Table III can be used to compare local (near the solution) rates of convergence—FCRM fares somewhat better than EM in this regard. The per iteration work required to execute fuzzy c-regression models depends on $E_{ik}(\beta_i)$. In this example EM iteration has the additional computational disadvantage of requiring 400 exponential function evaluations per iteration.

Also given in Table III are the average resubstitution error rates (as percentages) that are based on defuzzification or deprobabilization of the terminal partition matrices found by each algorithm. To calculate these error rates, each terminal partition matrix was converted to a hard partition matrix using the rule of maximum membership or maximum probability; e.g., if $U_{16} = 0.32$ and $U_{26} = 0.68$, then $(x_6,\, y_6)$ was assigned to the second cluster, and then this was counted as a labeling error if and only if $(x_6,\, y_6)$ was actually generated using the first model. Note that error rates for the two methods are practically the same in all three cases.

Finally, Table III shows average estimates of the "true but unknown" parameters $\{\alpha, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}\}$ obtained using fuzzy c-regression models and the EM algorithm in each case as $\{\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{21}, \hat{\beta}_{22}.\}$ Each row in Table III that exhibits such an average is followed by the corresponding average squared error, e.g., $(\alpha - \hat{\alpha})^2$ between the true value and its

TABLE V
DESCRIPTION OF DATA GENERATED FROM THE QUADRATIC MODELSy $= \beta_{i1} + \beta_{i2}\mathbf{x} + \beta_{i3}\mathbf{x}^2$

Data Set	n	x Values Interval	β_1	β_2
A	46	[5 ,27.5]	$\beta_{1A} = (21, -2, 0.0625)$	$\beta_{2A} = (-5, 2, -0.0625)$
B	28	[9,22.5]	$\beta_{1B} = (21, -2, 0.0625)$	$\beta_{2B} = (-5, 2, -0.0625)$
C	30	[9 , 23.5]	$\beta_{1C} = (18, -1, 0.03125)$	$\beta_{2C} = (-2, 1, -0.03125)$
D	46	[10.5,21.75]	$\beta_{1D} = (172, -26, 1)$	$\beta_{2D} = (364, -38, 1)$

TABLE VI
ITERATIONS REQUIRED FOR TERMINATION IN EXAMPLE 2

Initial Parameters (the dashed lines in Figs. 6-8)			Iterations of FCRM to Termination			
Init.	Initial β_1	Initial β_2	A	B	C	D
I_1	$\beta_{1,0} = (-19, 2, 0)$	$\beta_{2,0} = (-31, 2, 0)$	7 ($\surd$)	6 ($\surd$)	6 ($\surd$)	10 ($\surd$)
I_2	$\beta_{1,0} = (9, 0, 0)$	$\beta_{2,0} = (7, 0, 0)$	16 (X)	5 ($\surd$)	4 ($\surd$)	30 (X)
I_3	$\beta_{1,0} = (-8000, 1000, 0)$	$\beta_{2,0} = (-24000, 1000, 0)$	20 (X)	14 ($\surd$)	6 ($\surd$)	11 ($\surd$)

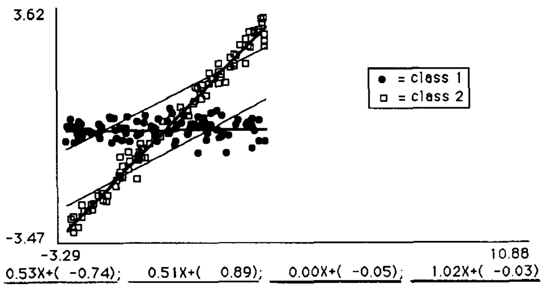


Fig. 5. A case 1 sample where both a global (heavy lines) and local (thin lines) minimizer are found by FCRM.

estimate over 25 sets of samples in each case. Note that the estimates are both close to each other, as well as to the true values.

Sensitivity of each method to initialization was also studied, and is summarized in Table IV. Fuzzy c-regression models and EM were both fairly insensitive to the initial guess as long as a "reasonably good" initialization $U^{(0)}$ was used. There was a difference, however, when extremely poor initializations were used. To examine this, both algorithms were started using an initial partition that grouped the first 100 data drawn into cluster 1, and the second 100 data drawn into cluster 2.

Table IV lists the number of times (out of 25 trials) that each algorithm terminated at extrema different from those obtained using the exact initialization. Notice that EM demonstrated complete insensitivity to all initializations used, while the fuzzy c-regression models algorithm produced additional extrema in a nonnegligible number of cases.

A typical example of FCRM termination at an undesirable extremum is depicted in Fig. 5. We note that undesirable extrema were always local in the cases examined. Apparently the global minimum of the fuzzy c-regression models function always occurred for parameter values near the true values, but the algorithm was sometimes trapped at a local solution. It is fairly well established that fuzzy models based on functionals such as (9) involving mth powers of the U_{ik}'s can sometimes avoid local trap states by using large values of m in the initial stages of iteration. We made no attempt to study this possibility here, but mention it as a promising strategy for improvement of FCRM in this regard.

Example 2

This example illustrates the use of FCRM to fit $c = 2$ quadratic regression models. The two quadratic models considered are of the form

$$y = \beta_{11} + \beta_{12}x + \beta_{13}x^2, \tag{22a}$$

and

$$y = \beta_{21} + \beta_{22}x + \beta_{23}x^2. \tag{22b}$$

The function that we attempt to iteratively minimize in the following experiments is given by (9) with $m = 2$ and $E_{ik}(\beta_i) = (y_k - \beta_{i1} - \beta_{i2}x_k - \beta_{i3}x_k^2)^2$. Four sets of data, named A, B, C and D, as specified in Table V, were generated for the tests. Each of the four data sets was generated by computing y from 22(a) or 22(b) at $n/2$ fixed, equally spaced x values across the interval given in column 3 of Table V. This resulted in sets of n points (which we pretend are unlabeled), half of which were generated from each of the two quadratics specified by the parameters in columns 4 and 5 of Table V. These four data sets are each plotted in Figs. 6, 7, and 8.

FCRM iterations seeking three pairs of quadratic models were initialized at pairs of *lines*. The quadratic parameters for these three initializations, named I_1, I_2, and I_3, are shown as columns 1 and 2 of Table VI. Coefficients of all x^2 terms for each initialization are zero, resulting in the initializing dashed lines shown in Figs. 6, 7 and 8, respectively.

Iteration for each of the three initializations was stopped as soon as the maximum change in the absolute value of successive pairs of estimates of the six parameter values for that model was found to be less than or equal to $\varepsilon = 0.00001$. Table VI lists the numbers of iterations of FCRM needed to achieve this termination criterion for each of data sets A-D. In Table VI the symbol ($\surd$) indicates that FCRM terminated at (a close numerical approximation to) the parameters of the true model, while the symbol X means that FCRM terminated at

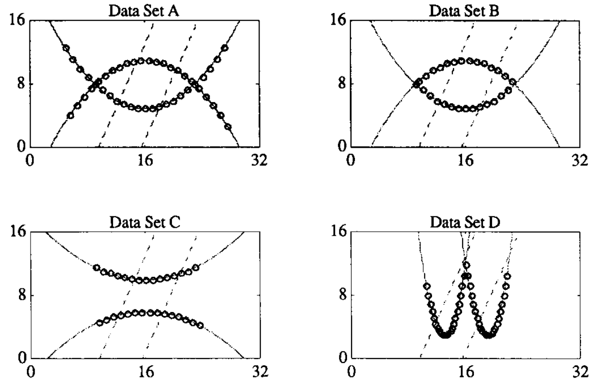


Fig. 6. Initial (dashed lines) and terminal models (solid arcs) for initialization I_1.

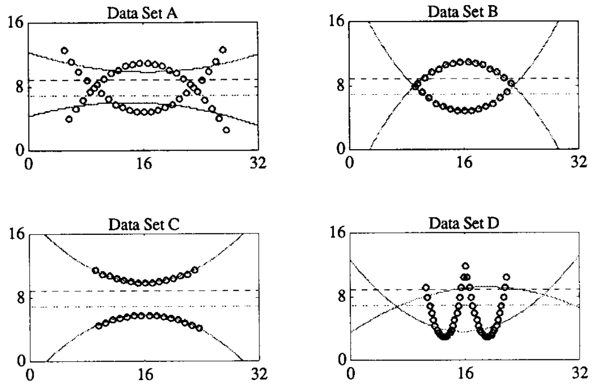


Fig. 7. Initial (dashed lines) and terminal models (solid arcs) for initialization I_2.

incorrect estimates. FCRM terminated at the correct solution for all initializations with data sets B and C, failed once using I_2 with data set D, and suffered two failures with I_2 and I_3 on data set A. To interpret these results more carefully, we turn to Fig. 6-8.

Fig. 6 shows data sets A–D and graphs of the initial (dashed lines) and terminal regression models of FCRM found with initialization I_1. As indicated in row 1 of Table VI, this initialization led to successful termination at the true values of the generating quadratics very rapidly (6–10 iterations) for all four data sets. Observe that for this case, the initializing lines were neither horizontal nor vertical—they were inclined to the axes of symmetry of the data in every case. The fits to the data are in these four cases quite good (in fact, they were accurate to essentially machine precision). We remind you that the data sets are fed to FCRM unlabeled, but we have cheated a little, knowing ahead of time that $c = 2$.

Fig. 7 depicts the same information as Fig. 6 for the second initialization of Table VI. This initialization, the horizontal line pair shown in Fig. 7, leads FCRM astray twice. Specifically, FCRM terminates incorrectly at the solid arcs shown for data sets A and D. Since A has both horizontal and vertical symmetry, while D has only vertical symmetry, we probably should not attribute both failures to the same cause (which is unknown to us at this time).

Figure 8 shows the analogous outputs of FCRM for the third initialization, which is the pair of nearly vertical dashed lines in each subfigure. The correct models are found for data sets B, C, and D, but FCRM fails again on data set A. Comparing Figs. 7 and 8, data set A appears to be the most difficult case because of the "X-ed" points to the left and right of the intersections of the two generating quadratics. We did not attempt to rectify these outputs for this note, but suspect that this can be accomplished with a pruning or cleanup procedure guided by the membership values for each data point in the terminal partition U generated by FCRM.

In this example FCRM detects and characterizes the quadratic models generating these four data sets correctly in nine of 12 attempts. The artificial data chosen for this example are not particularly interesting except as a means for illustrating that FCRM works nicely with data generated by nonlinear switching regression models.

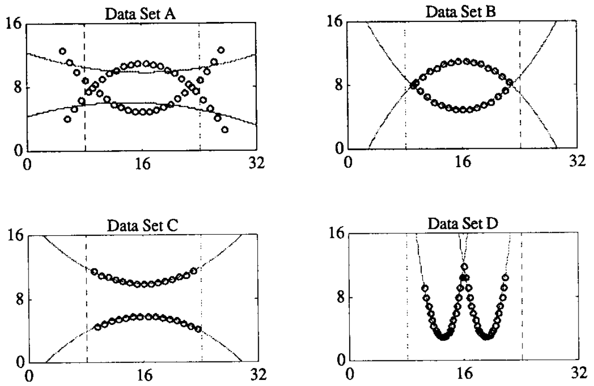


Fig. 8. Initial (dashed lines) and terminal models (solid arcs) for initialization I_3.

IV. Discussion

For Example 1, FCRM required fewer iterations than EM, but sometimes terminated at undesirable estimates when poor initializations were used. The quality of FCRM estimates compared well with the (asymptotically efficient) ML estimates in cases 1 and 2 when $\alpha = 0.5$ and $\sigma_1 = \sigma_2$. Case 3 indicates that the maximum likelihood estimates may be better able to handle cases where σ_1 is significantly different from σ_2, which is not surprising since estimates of the standard deviations are calculated and used during each EM iteration. Perhaps the most useful conclusion one may draw from the values exhibited in Table III is that, on balance, the two methods produced very similar results. Differences in the various values and squared errors of the estimates are deemed insignificant in terms of affording a comparative advantage to either technique.

Allowing $m = 1$ in (9) would give a family of hard (or crisp) c-regression models functions. The corresponding family of HCRM algorithms is identical to the fuzzy version stated except that (11) is replaced by:

Step 3′. Update $U^{(r)} \rightarrow U^{(r+1)} \in M_{cn}$, with $E_{ik} = E_{ik}(\beta_i^{(r)})$, to satisfy

$$U_{ik} = 0 \text{ if } E_{ik} > E_{jk} \text{ for some } j, \\ \text{and } U_{ik} \in \{0, 1\} \text{ with } (U_{1k} + \cdots + U_{ck}) = 1. \tag{23}$$

A problem this technique would share with other hard clustering approaches occurs when the clusters have significant overlap. Hard partitionings force each datum $(\mathbf{x}_k, \mathbf{y}_k)$ to be totally classified into exactly one of the c clusters, while fuzzy partitionings offer more representational power by differentiating, through a range of values of the membership functions, between points that are clearly in a particular cluster and those that are not. This advantage of fuzzy clustering is well known [11].

Readers familiar with fuzzy clustering know that there are at least *three* other approaches to fitting linear clusters such as those in Figs. 2-4. All three approaches, as well as the one advocated here, are driven by optimization of a generalized fuzzy c-prototypes functional, say $J(U, P; S) = \sum_{k=1}^{n} \sum_{i=1}^{c} (U_{ik})^m D_i^2(\mathbf{z}_k, P_i)$, where $D_i(\mathbf{z}_k, P_i)$ is a measure of similarity (or dissimilarity) between datum $\mathbf{z}_k$ and *prototype* P_i. In this paper, $(D_i)^2 = E_i$ and the "prototypes" that are used to best fit the data are regression models, $P_i = f_i(\mathbf{x}; \beta_i)$, where $z = (\mathbf{x}, \mathbf{y})$. Previous approaches to fitting linear clusters include those proposed by Gustafson and Kessel [23], wherein each fitting prototype is a vector, $P_i = \mathbf{v}_i$, and the linear shapes are matched through distortions of inner product norms (individual to each cluster) induced by positive definite weight matrices, $D_i(\mathbf{z}_k, P_i) = \|\mathbf{z}_k - \mathbf{v}_i\|_{A_i}$; Bezdek *et al.* [24], [25], where the fitting prototypes are either straight lines, $P_i = \mathbf{v}_i + t\mathbf{d}_i$ in $\mathbf{R}^{s+t}$ and $D_i(\mathbf{z}_k, P_i)$ is the orthogonal distance from $\mathbf{z}_k$ to P_i, or more generally, prototypes that are convex combinations of points and lines; and Dave [26], where each fitting prototype P_i is a hyperquadric shell, and the distance from $\mathbf{z}_k$ to P_i is defined by a well-considered geometric rationale.

We emphasize that FCRM differs from these methods in *three* important ways. First, our "prototypes" are not geometric objects—they are functions. Second, since the prototypes are functions, FCRM is not explicitly designed to find linear clusters. Because the models are functions, FCRM can be used, for example, when the regression models are quadratic (as in Example 2), cubic, or even transcendental functions. Thus, data whose functional dependency is much more complicated than linear can (in principle at least) be easily accommodated by FCRM. And finally, FCRM explicitly recognizes functional dependency between grouped subsets of coordinates in the data, whereas none of the previous methods do. These are major differences between FCRM and all previous approaches. Example 2 was included here to illustrate that FCRM is not limited to fitting linear structures.

Interesting questions for future research include:

(i) A robustness study and thorough investigation of the relative merits of different choices for the error measures $\{E_{ik}(\beta_i)\}$ as well as the appropriate parametric families $\{\mathbf{f}_i(\mathbf{x}; \beta_i)\}$ to use in the model.

(ii) A numerical study comparing empirical convergence properties of FCRM with fuzzy c-varieties from [11] on problems solvable by both techniques.

(iii) Studying the use of FCRM for fitting mixed data with nonlinear dependencies.

(iv) Extension of FCRM to the case of unconstrained labels, by adapting the possibilistic clustering approach of Krishnapuram and Keller [15] to the current problem. This scheme might be less prone to termination at local minimizers.

(v) Determining the number of clusters in a population, which in this case becomes the number of models that are needed.

(vi) FCRM suffers from several problems characteristic of all calculus-based optimization methods, viz., guessing good initializations and avoiding local trap states. A study of these problems is warranted.

Our results establish FCRM as a promising technique for switching regression parameter estimation and clustering. We observe particularly that the FCRM approach is motivated and implemented without recourse to particular statistical assumptions on ε_i; it is therefore hoped that fuzzy c-regression models will perform satisfactorily in many different cases. A large-scale simulation will provide more insight into the robustness of this new approach. In summary, FCRM converges very rapidly and produces high-quality estimates often enough to justify further study. We hope to make some of the above issues the basis of a future investigation.

References

[1] R. D. De Veaux, "Mixtures of linear regressions," *Computational Statistics and Data Analysis*, vol. 8, pp. 227–245, 1989.

[2] M. Aitkin and G. T. Wilson, "Mixture models, outliers, and the EM algorithm," *Technometrics*, vol. 22, no. 3, pp. 325–331, 1980.

[3] J. E. Dennis, Jr., "Algorithms for nonlinear fitting," in *Proc. NATO Advanced Research Symposium*, (Cambridge University), 1981.

[4] D. S. Hamermesh, "Wage bargains, threshold effects, and the phillips curve," *Quarterly Journal of Economics*, vol. 84, pp. 501–517, 1970.

[5] D. W. Hosmer, Jr., "Maximum likelihood estimates of the parameters of a mixture of two regression lines," *Communications in Statistics*, vol. 3, no. 10, pp. 995–1005, 1974.

[6] N. M. Kiefer, "Discrete parameter variation: efficient estimation of a switching regression model," *Econometrica*, vol. 46, no. 2, pp. 427–434, 1978.

[7] R. E. Quandt, "A new approach to estimating switching regressions," *J. Amer. Statist. Ass.*, vol. 67, pp. 306–310, 1972.

[8] R. E. Quandt and J. B. Ramsey, "Estimating mixtures of normal distributions and switching regressions," (with discussion), *J. Amer. Statist. Ass.*, vol. 73, pp. 730–752, 1978.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[10] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[11] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[12] J. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[13] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[14] E. H. Ruspini, "A new approach to clustering," *Inform. Control*, vol. 15, pp. 22–32, 1969.

[15] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, pp. 98–110, May 1993.

[16] J. C. Bezdek and J. C. Dunn, "Optimal fuzzy partitions: a heuristic for estimating the parameters of a mixture of normal distributions," *IEEE Trans. Comput.*, vol. 24, pp. 835–838, 1975.

[17] J. C. Bezdek, R. J. Hathaway, R. E. Howard, C. A. Wilson, and M. P. Windham, "Local convergence analysis of a grouped variable version of coordinate descent," *Journal of Optimization Theory and Applications*, vol. 54, no. 3, pp. 471–477, 1987.

[18] R. J. Hathaway and J. C. Bezdek, "Grouped coordinate minimization using Newton's method for inexact minimization in one vector coordinate," *Journal of Optimization Theory and Applications* vol. 71, no. 3, pp. 503-516, 1991.

[19] W. Zangwill, *Non-Linear Programming: A Unified Approach.* Englewood Cliffs, NJ: Prentice Hall, 1969.

[20] J. C. Bezdek and R. J. Hathaway, "Numerical convergence and interpretation of the fuzzy c-shells clustering algorithm," *IEEE Trans Neural Networks*, vol. 3, pp. 787–793, Sept. 1992.

[21] C. Bays and S. D. Durham, "Improving a poor random number generator," *TOMS* , vol. 2, pp. 59–64, 1976.

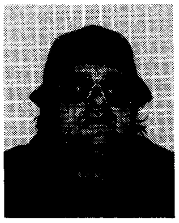
[22] G. E. P. Box and M. E. Muller, "A Note on the generation of random normal deviates," *Ann. Math. Statist.*, vol. 29, pp. 610–611, 1958.

[23] E. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE CDC* (San Diego, CA), 1979, pp. 761–766.

[24] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure: I. Linear structure: fuzzy c-lines," *SIAM J. Appl. Math.*, vol. 40, no. 2, pp. 339–357, 1981.

[25] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure: II. Fuzzy c-varieties and convex combinations thereof," *SIAM J. Appl. Math*, vol. 40, no. 2, pp. 358–372, 1981.

[26] R. N. Dave, "Use of the Adaptive fuzzy clustering algorithm to detect lines in digital images," in *Proc. SPIE Conf. on Intelligent Robots and Computer Vision VIII*, 1989, pp. 600–611.



Rick (PeachFuzz) Hathaway received the B.S. degree in applied mathematics from the University of Georgia in 1979 and the Ph.D. degree in mathematical sciences from Rice University in 1983.

He is an Associate Professor in the Mathematics and Computer Science Department of Georgia Southern University. His research interests include pattern recognition, statistical computing, and nonlinear optimization.

Jim (NoFuzz) Bezdek (M'80–SM'90–F'92) received the B.S. degree in civil engineering from the University of Nevada (Reno) in 1969 and the Ph.D. in applied mathematics from Cornell University in 1973.

He is a Professor in the Computer Science Department at the University of West Florida. His research interests include pattern recognition, image processing, computational neural networks, and medical applications.