BIT 22 (1982), 140–149

CHOOSING THE r-DIMENSION FOR THE FCV FAMILY OF CLUSTERING ALGORITHMS

ROBERT W. GUNDERSON

Institute of Informatics, University of Oslo, Blindern, Oslo 3, Norge

Department of Mathematics, Utah State University, Logan, Utah 84322

Abstract.

A strategy is given for selecting the dimension r of the linear variety which is used to define the criterion functional J_{vrm} and which determines the shape of the data clusters detected by the corresponding c-Varieties (FCV) clustering algorithms.

1. Introduction.

One of the questions not answered in [1] was that of a criterion for selecting r, the dimension of the linear variety V_r used in the definition of the criterion function J_{vrm} and which ultimately determines the shape of the clusters detected by the FCV algorithms. A poor choice for r can lead to disappointing or misleading results since, in such cases, one effectively attempts to impose upon the data a structure which does not exist. Ideally then, one would like to make the choice of r data dependent. However, this objective is complicated by the fact that a limited knowledge about the structure of the data is usually the reason for a cluster analysis in the first place.

In this note a modified family of FCV algorithms is suggested in which r is treated as a parameter to be determined at each step of the FCV iterative procedure. The criterion for selecting r is thus based upon structural properties *encountered* in the data. By adapting to a data dependent linear variety the modified algorithms thereby tend to *seek out cluster shapes* and reduce the risk of arbitrarily imposing a non-representative structure.

As in [1], we shall assume that all of the clusters in the data set are of the same general shape; that is, satisfactorily modelled by a prototype defined by a linear variety of common dimension r. An investigation using a similar approach for detecting a *mixture* of different cluster shapes is in progress and discussed briefly in a remark at the conclusion of this paper. A more detailed discussion of this important case will be left to a subsequent note.

Received May 22, 1981. Revised April 13, 1982.

2. The FCV family of clustering algorithms.

Following the notation and definitions of [1], let $X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$ be a finite sample of an unlabeled data set, $X \subset \mathbb{R}^s$, and let U denote a real $c \times n$ matrix with $2 \leqq c < n$ and elements u_{ik} satisfying

$$u_{ik} \in [0, 1] \qquad \forall\, i, k \tag{2.1a}$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall\, k \tag{2.1b}$$

$$0 < \sum_{k=1}^{n} u_{ik} \qquad \forall\, i. \tag{2.1c}$$

By letting the c characteristic functions $u_i : X \to [0, 1]$ be defined according to

$$u_i(x_k) = u_{ik} \qquad \forall\, i, k \tag{2.2}$$

the matrix U can be interpreted as establishing a (fuzzy) c-partition over the sample data set X. The details are given in [1].

Now, for each $i = 1, 2, \ldots, c$ let

$$V_i(\bar{v}_i; \bar{d}_{i1}, \bar{d}_{i2}, \ldots, \bar{d}_{ir}) = \{\bar{y} \in \mathbb{R}^s | \bar{y} = \bar{v}_i + \sum_{j=1}^{r} t_j \bar{d}_{ij}; t_j \in \mathbb{R}\} \tag{2.3}$$

denote a linear variety of dimension $r, 0 \leqq r < s$, in $\mathbb{R}^s$ through $\bar{v}_i$ and spanned by an orthonormal set of vectors $\{\bar{d}_{i1}, \bar{d}_{i2}, \ldots, \bar{d}_{ir}\}$. Define the orthogonal distance of a sample vector $\bar{x}_k \in \mathbb{R}^s$ to the linear variety V_i by

$$D_{ik} = D(\bar{x}_k, V_i) = (\|\bar{x}_k - \bar{v}_i\|_A^2 - \sum_{j=1}^{r} (\langle \bar{x}_k - \bar{v}_i, \bar{d}_{ij} \rangle_A)^2)^{\frac{1}{2}} \tag{2.4}$$

where

$$\langle \bar{x}, \bar{y} \rangle_A = \bar{x}^T A \bar{y} \tag{2.5}$$

for the positive definite $s \times s$ matrix A and

$$\|\bar{x}\|_A^2 = \langle \bar{x}, \bar{x} \rangle_A. \tag{2.6}$$

The c-Varieties (FCV) family of algorithms presented in [1] follow from the necessary conditions of the following two theorems for minimizing the generalized weighted sum-of-squared-error criterion functional

$$J_{vrm}(U, \bar{V}) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (D_{ik})^2 \tag{2.7}$$

where $\bar{V} = \{V_1, V_2, \ldots, V_c\}$ and the minimization is carried out for fixed $1 < m < \infty$ over $M_{fc} \times \mathbb{R}^{cs} \times (\mathbb{R}^{cs})^r$ and where M_{fc} denotes the family of partition matrices satisfying conditions 2.1 a), b) and c):

THEOREM 1. *Let $\bar{V} = \hat{V} \in \mathbb{R}^{cs} \times (\mathbb{R}^{cs})^r$ be fixed and assume $1 < m < \infty$ and $\hat{D}_{ik} > 0$ $\forall\, i, k$. Then, $\hat{U} \in M_{fc}$ is a strict local minimum of $\mathcal{O}(U) = J_{vrm}(U, \hat{V})$ if and only if*

$$\hat{u}_{ik} = 1/\sum_{j=1}^{c} (\hat{D}_{ik}/\hat{D}_{jk})^{2/(m-1)}, \qquad \forall\, i, k. \tag{2.8}$$

THEOREM 2. *Let $\hat{U} \in M_{fc}$ be fixed and $1 < m < \infty$. Then $\hat{V} \in \mathbb{R}^{cs} \times (\mathbb{R}^{cs})^r$ is a local minimum of $\psi(\bar{V}) = J_{vrm}(\hat{U}, \bar{V})$ only if for*

$$\bar{v}_i = \sum_{k=1}^{n} (\hat{u}_{ik})^m \bar{x}_k / \sum_{k=1}^{n} (\hat{u}_{ik})^m, \qquad \forall\, i \tag{2.9a}$$

we set

$$\hat{d}_{ij} = A^{-\frac{1}{2}} \hat{y}_{ij} \qquad (j = 1, 2, \ldots, r) \tag{2.9b}$$

where $\hat{y}_{ij}$ is the unit eigenvector corresponding to the jth largest eigenvalue of the matrix $A^{\frac{1}{2}} S_i A^{\frac{1}{2}}$ and where

$$S_i = \sum_{k=1}^{n} (\hat{u}_{ik})^m (\bar{x}_k - \hat{v}_i)(\bar{x}_k - \hat{v}_i)^T, \tag{2.9c}$$

is the within-cluster scatter of the ith (fuzzy) cluster.

REMARK 1. Conditions (2.9b) could be written in the equivalent form

(2.9b′) $\hat{d}_{ij}$ is the unit eigenvalue corresponding to the jth largest eigenvalue of the matrix $S_i A$.

After taking into account the singular case where $\hat{D}_{ik} = 0$ for at least one pair (i, k) in Theorem 1, the results above provide a Picard iteration procedure which is shown in [1] strictly to descend to a local minimum of J_{vrm}, at least in the nonsingular case and for $1 < m < \infty$:

I) For fixed $\hat{V} = \{\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_c\}$ let

$$N_k = \sum_{i=1}^{c} n_i \qquad \text{where } n_i = \begin{cases} 1, & \hat{D}_{ik} = 0 \\ 0, & \hat{D}_{ik} \neq 0 \end{cases} \qquad k = (1, 2, \ldots, n)$$

i) If $N_k = 0$ $(k = 1, 2, \ldots, n)$ then set

$$\hat{u}_{ik} = 1/\sum_{j=1}^{c} (\hat{D}_{ik}/\hat{D}_{jk})^{2/(m-1)} \tag{2.10}$$

ii) If $N_k > 0$ $(k = 1, 2, \ldots, n)$ then set

$$u_{ik} = \begin{cases} 0, & \hat{D}_{ik} \neq 0 \\ 1/N_k, & \hat{D}_{ik} = 0. \end{cases} \tag{2.11}$$

II) If $N_k > 0$, $k = 1, 2, \ldots, n$ then STOP. $(\hat{U}, \hat{V})$ from I is a (possibly degenerated) minimizing solution. (U is degenerate if there exists a j such that $\sum_{k=1}^{n} u_{jk} = 0$.)

III) For fixed $\hat{U} = \{\hat{u}_{ik}\}$ $(i = 1, \ldots, c; k = 1, 2, \ldots, n)$, compute $\hat{V}$ from (2.9 a, b and c).

Starting with either $\hat{U}$ or $\hat{V}$ loop through I, II, III until a sufficiently good approximation to a minimizing partition is obtained.

REMARK 2. Notice that the computation of the $\hat{u}_{ik}$ from (2.8) is independent of the method used to determine the distances $\hat{D}_{ik}$. Theorem I only assumes that the $\hat{D}_{ik}$ are supplied. Similarly, calculation of the centers, $\bar{v}_i$, is dependent only upon the $\hat{u}_{ik}$ and therefore independent of the method used to generate the $\hat{D}_{ik}$. Since the scatter matrices S_i are dependent only upon the $\hat{u}_{ik}$ and $\bar{v}_i$ in (2.9c), no decision has yet been necessary regarding the dimension r of the linear varieties defining each individual cluster, in looping through the iterative procedure.

3. A scatter criterion for selecting the dimension r.

The first part of the proof to theorem 2 in [1] shows that if a solution V exists then the $\hat{v}_i$ can be chosen according to (2.9a) for every $i = 1, 2, \ldots, c$. The second part, concerning (2.9b), proceeds by noting that J_{vrm} will be minimized if and only if the c individual terms

$$\psi_i(V_i) = \sum_{k=1}^{n} (u_{ik})^m (D_{ik})^2 \tag{3.1}$$

are minimized, $i = 1, 2, \ldots, c$. Minimizing (3.1) over the spanning vectors $\bar{d}_{ij}$ for fixed $\hat{v}_i$ and $j = 1, 2, \ldots, c$ is equivalent to maximizing

$$\sum_{k=1}^{n} (\hat{u}_{ik})^m \sum_{j=1}^{r} \langle x_k - v_i, \bar{d}_{ij} \rangle_A^2 \tag{3.2}$$

over $\bar{d}_{ij}$ and (3.2) can be rewritten as

$$\sum_{j=1}^{r} \bar{d}_{ij}^T A \Big(\sum_{k=1}^{n} (\hat{u}_{ik})^m (\bar{x}_k - \hat{v}_i)(\bar{x}_k - \hat{v}_i)^T \Big) A \bar{d}_{ij}. \tag{3.3}$$

Thus we are led to consider the problem

$$\text{maximize} \quad \sum_{j=1}^{r} \bar{d}_{ij}^T A S_i A \bar{d}_{ij} \tag{3.4}$$

over all orthonormal subsets $\{\bar{d}_{i1}, \bar{d}_{i2}, \ldots, \bar{d}_{ir}\}$ for each $i = 1, 2, \ldots, c$, i.e.

$$\text{maximize}_{\bar{z}_j \neq 0} \sum_{j=1}^{r} \frac{\bar{z}_j^T A S_i A \bar{z}_j}{\bar{z}_j^T A \bar{z}_j} \tag{3.5}$$

over all orthogonal subsets $\{\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_r\}$.

The solution to this problem follows from a theorem found in [2, p. 322]. According to that theorem, if λ_1 is the largest of the s real eigenvalues of the matrix $S_i A$, then

$$\max_{\bar{z} \neq 0} \frac{\bar{z}^T A S_i A \bar{z}}{\bar{z}^T A \bar{z}} = \lambda_1 \tag{3.6}$$

and this maximum is assumed only for eigenvectors form the eigenspace of $S_i A$ corresponding to λ_1. Further, the jth largest eigenvalue

$$\lambda_s \leqq \ldots \leqq \lambda_{j+1} \leqq \lambda_j \leqq \lambda_{j-1} \leqq \ldots \leqq \lambda_1 \tag{3.7}$$

is the maximum of the same ratio, with $\bar{z}$ constrained to lie in the subspace orthogonal to that spanned by the eigenvectors corresponding to λ_{j-1} through λ_1, and

$$\max_{\bar{z} \neq 0} \frac{\bar{z}^T A S_i A \bar{z}}{\bar{z}^T A \bar{z}} = \lambda_j \tag{3.8}$$

is obtained only for those eigenvectors from the eigenspace of $S_i A$ corresponding to λ_j. Conditions (2.9 b, c) follow in the form given by Remark 1.

Adopting the terminology of multiple discriminant analysis (cf. [1]) eq. (3.1) can be interpreted as the total scatter of the data X onto the subspace of $\mathbb{R}^s$ spanned by the set of vectors $\{A\bar{d}_{i1}, A\bar{d}_{i2}, \ldots, A\bar{d}_{ir}\}$, weighted by the mth power of the membership values of the data vectors $\bar{x}_k$ in the ith partition of X. Roughly speaking the linear variety of dimension r and through v_i which minimizes ψ_i of (3.1) is the one on which there is maximum scatter and equation (3.8) supplies a device to measure the relative amounts of scatter in the maximizing directions of the orthonormal vectors $\bar{d}_{ij}$, namely by viewing the ratios of the eigenvalues of (3.7). (One may also note the application of the FCV algorithms to principal components analysis.)

In Remark 2 it was noted that the dimension r could be assigned at the beginning of each loop through the FCV algorithms. The preceding discussion suggests that

the assigned value might be based upon the shape of the clusters, using the scatter information provided by each of the c within-cluster scatter matrices S_i. The following implementation provides an "adaptive" FCV family of algorithms which has provided good results on test sets of data, where the data are known to consist of a fixed number of clusters of uniform, i.e. the same general, shape.

1. Start the FCV algorithm with $r = 0$.
2. Replace step III of the algorithm by:
   i) Compute cluster centers $\hat{v}_i$ from (2.9a) ($i = 1, 2, \ldots, c$);
   ii) Compute the within-cluster scatter matrices S_i from (2.9c) ($i = 1, 2, \ldots, c$);
   iii) Compute eigenvalues $\lambda_{is} \leqq \ldots \leqq \lambda_{i1}$ and corresponding eigenvectors $\bar{d}_{i1}, \ldots, \bar{d}_{is}$ ($i = 1, 2, \ldots, c$);
   iv) Set $r = k \qquad k = 1, 2, \ldots, s-1$

$$\text{if } \varepsilon_0 \leqq \frac{\hat{\lambda}_{i,k+1}}{\hat{\lambda}_{ik}} < \varepsilon_p \qquad \begin{matrix} i = 1, 2, \ldots, c \\ p = 0, 1, \ldots, k-1 \end{matrix}$$

where

$\varepsilon_0 = 0$ and $\varepsilon_p \in (0, 1)$ for each $p > 0$: Otherwise, set $r = 0$.

   v) Return to step I.

This adaptive procedure was applied to the data shown in figures 1 and 2 ([3], p. 231). Figures 3 and 4 illustrate the consequences of an unfortunate selection for r in the usual case of the FCV algorithms and provides the motivation for this investigation. The results of figures 5 and 6 were obtained using the adaptive implementation of the FCV algorithms described above, with the same starting conditions used in both cases ($\varepsilon_1 = 0.25$, $r_0 = 0$, starting centers as shown, $m = \frac{3}{2}$).

REMARK 3. In the test cases run to date, a constant threshold value of $\varepsilon_p = 0.25$ for every $p = 1, 2, \ldots, s-1$ has resulted in satisfactory cluster results. The choice of these values is arbitrary and obviously will require some experimentation by the investigator on a given data set to arrive at values meaningful to that particular problem. This arbitrariness should not be viewed in an entirely negative fashion, however, for it provides the investigator with a tool to explore the shape of the clusters and thereby obtain a more complete view of the data structure.

REMARK 4. It was pointed out in part II of [1] that widely separated clusters of points falling on the same linear variety will not be identified as such by the FCV algorithms. The solution suggested there was to "penalize" the membership of such points according to their distance from the cluster-defining center v_i. Improved performance for the adaptive FCV algorithms can be obtained in the same way. Equation (2.4) then takes the form

$$D_{ik} = D(\bar{x}_k, \bar{v}_i) = (\|\bar{x}_k - \bar{v}_i\|_A^2 - \alpha \sum_{j=1}^{r} (\langle \bar{x}_k - \bar{v}_i, \bar{d}_{ij} \rangle_A)^2)^{\frac{1}{2}} \tag{2.4}$$

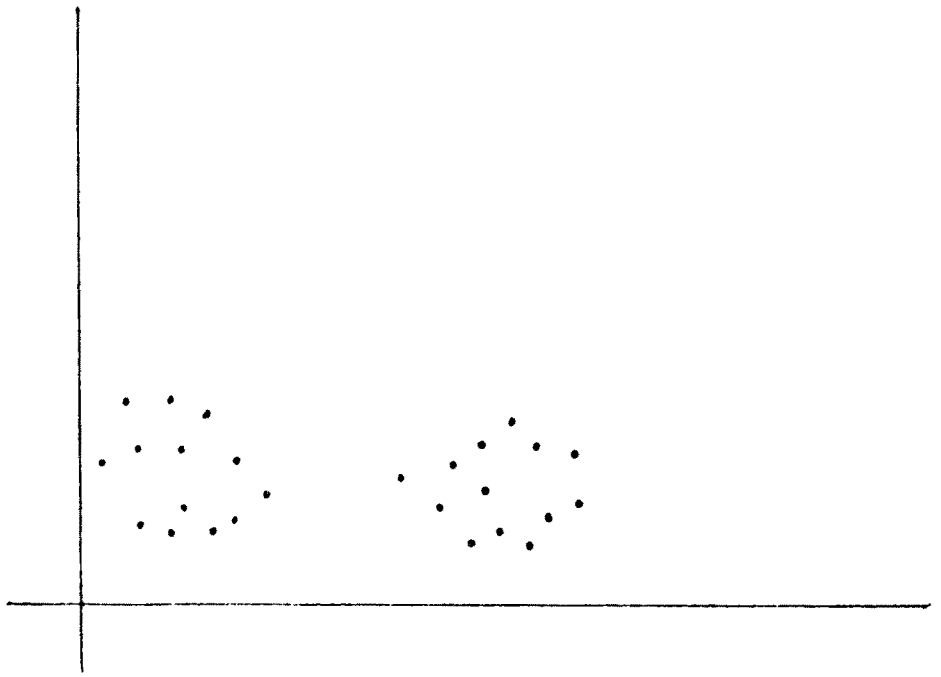

Fig. 1. Two round clusters.

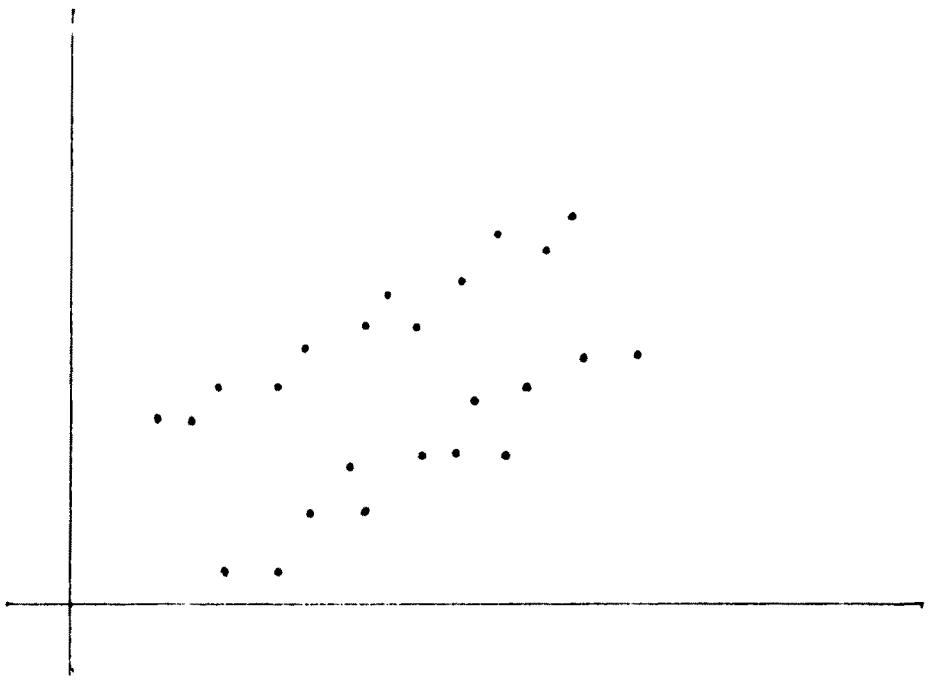

Fig. 2. Two linear clusters.

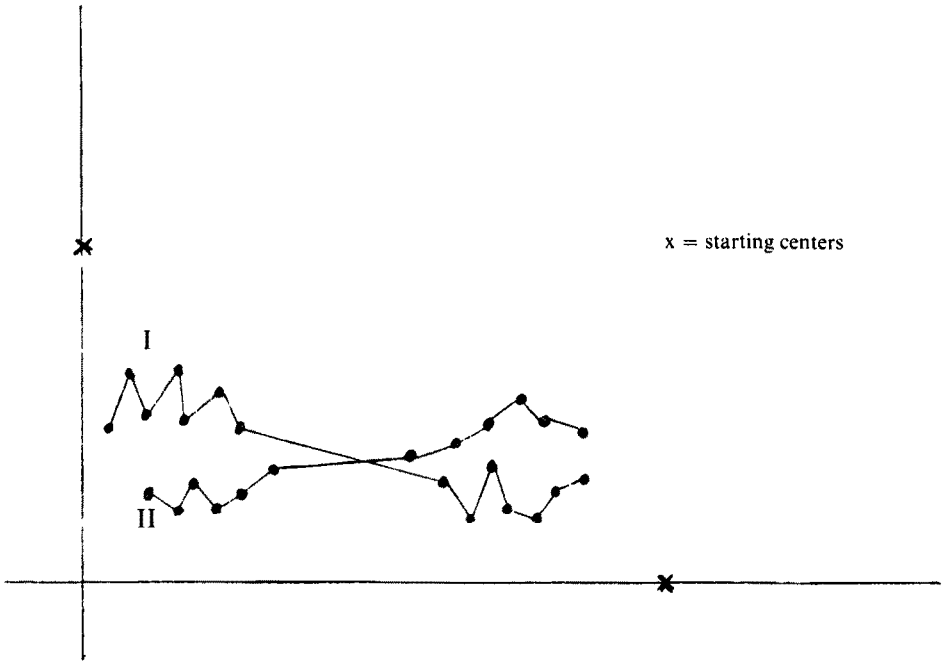

Fig. 3. Clusters detected with r = 1 (c-lines)

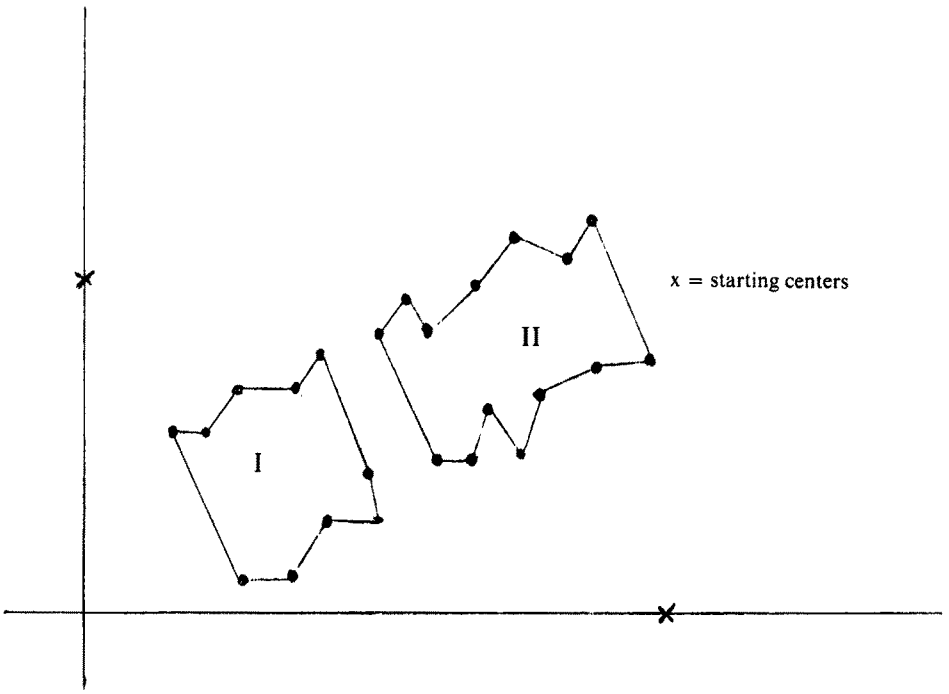

Fig. 4. Clusters detected with r = 0 (c-means).

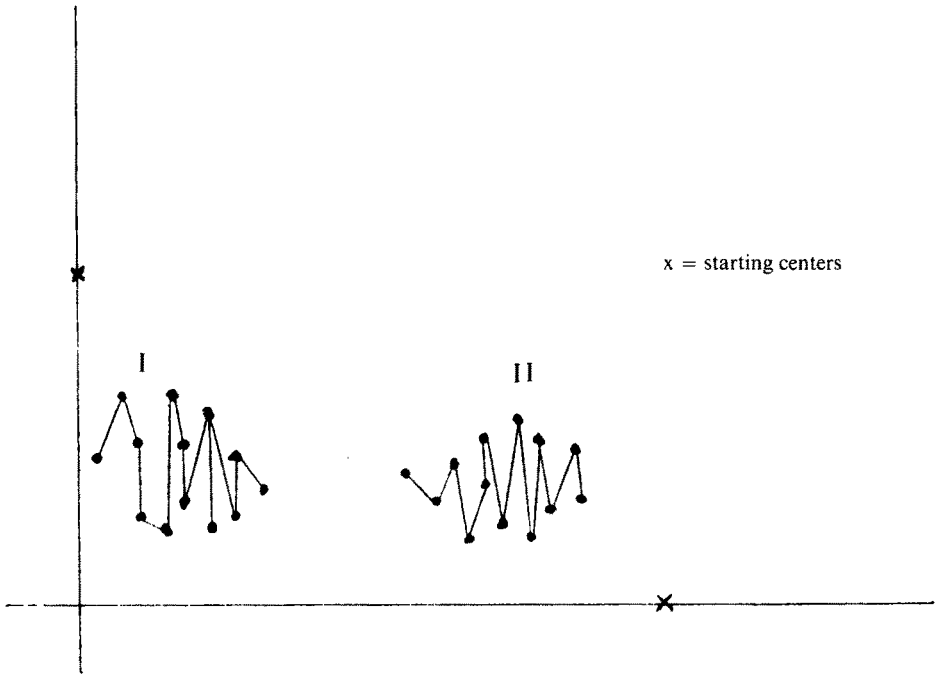

Fig. 5. Clusters detected by adaptive FCV algorithm.

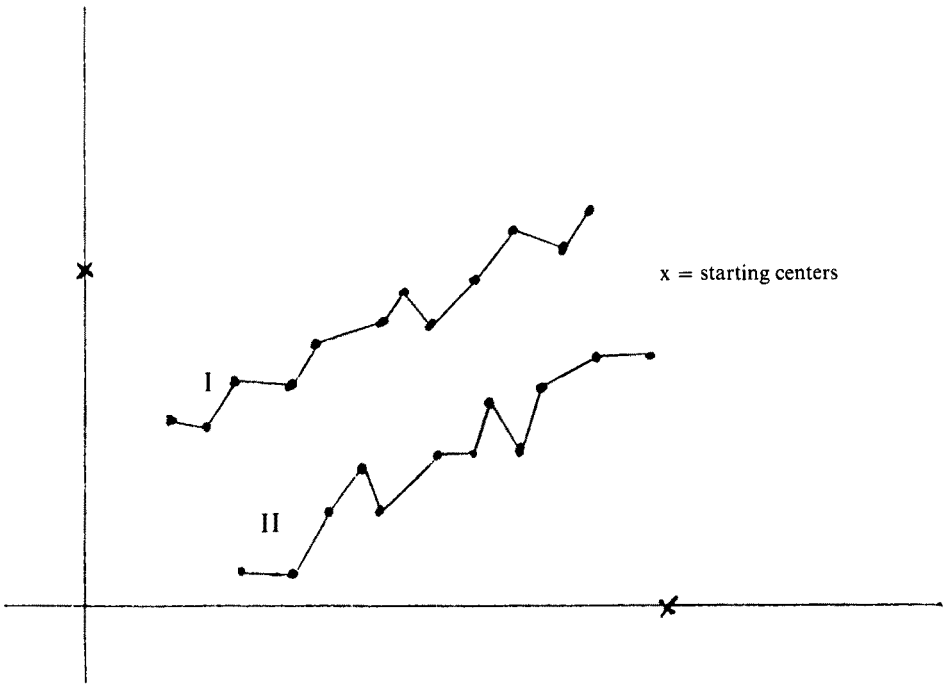

Fig. 6. Clusters detected by adaptive FCV algorithm

where $\alpha \in [0, 1]$. Following the suggestion in [1] of setting $\alpha \approx 0.9$ has continued to yield generally good results but, once again, a certain amount of experimenting with different values would be suggested for specific applications. The example problem of figures 5 and 6 was run using $\alpha = 1.0$.

REMARK 5. Once an equilibrium value for r has been achieved, one is back to the assumptions leading to the convergence results in [1]. Given the uniform cluster shapes in the test cases to date, this equilibrium has been reached after only a few iterations.

REMARK 6. The choice of r for cluster i can be made independent from the choice for cluster $j \neq i$, since the directions of maximum scatter are computed after v_i and v_j using eq. (2.9a). Thus it should be possible to follow basically the same approach to obtain an adaptive FCV algorithm which will seek out a *mixture* of clusters of different shapes within the data. An algorithm which will recognize a *mixture* of ball-like (point prototypes) and linear (line prototypes) clusters is providing very good results. Investigation of these algorithms is currently being focused upon convergence questions, which are complicated, among other factors, by the multiplication of possible stopping points for the iterative process.

REFERENCES

1. J. Bezdek, R. Gunderson, et al., *Detection and characterization of cluster substructure*, SIAM J. on Applied Mathematics, Vol. 40, No. 2, April 1981.
2. F. R. Gantmacher, *The Theory of Matrices*, Vol. I, Chelsea, 1959.
3. R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, Wiley-Interscience, New York, 1974.