On Cluster Validity for the Fuzzy c-Means Model

Nikhil R. Pal and James C. Bezdek

***Abstract*—Many functionals have been proposed for validation of partitions of object data produced by the fuzzy c-means (FCM) clustering algorithm. We examine the role a subtle but important parameter—the weighting exponent m of the FCM model—plays in determining the validity of FCM partitions. The functionals considered are the partition coefficient and entropy indexes of Bezdek, the Xie–Beni, and extended Xie–Beni indexes, and the Fukuyama–Sugeno index. Limit analysis indicates, and numerical experiments confirm, that the Fukuyama–Sugeno index is sensitive to both high and low values of m and may be unreliable because of this. Of the indexes tested, the Xie–Beni index provided the best response over a wide range of choices for the number of clusters, (2–10), and for m from 1.01–7. Finally, our calculations suggest that the best choice for m is probably in the interval [1.5, 2.5], whose mean and midpoint, $m = 2$, have often been the preferred choice for many users of FCM.**

I. Introduction

CLUSTERING in unlabeled data $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset \Re^p$ is the assignment of labels to the vectors in X and, hence, to the objects generating X. If the labels are hard (crisp), we hope they identify c natural subgroups in X. Clustering is also called unsupervised learning, with the word learning referring to learning the correct labels (and possibly vector prototypes or quantizers) for good subgroups in the data. c-partitions of X are sets of (cn) values $\{u_{ik}\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $U = [u_{ik}]$. There are three sets of partition matrices

$$M_{pcn} = \left\{ U \in \Re^{cn} \,|\, 0 \le u_{ik} \le 1 \quad \forall i, k; \right.$$
$$\left. \forall k,\, u_{ik} > 0 \quad \exists i; \quad 0 < \sum_{k=1}^{n} u_{ik} < n \quad \forall i \right\}; \quad \text{(1a)}$$

$$M_{fcn} = \left\{ U \in M_{pcn} \left| \sum_{i=1}^{c} u_{ik} = 1 \quad \forall k \right. \right\} \quad \text{(1b)}$$

$$M_{hcn} = \{U \in M_{fcn} \,|\, u_{ik} = 0 \quad \text{or} \quad 1 \quad \forall i \quad \text{and} \quad k\}. \quad \text{(1c)}$$

Equations (1) define, respectively, the sets of possibilistic, constrained fuzzy or probabilistic, and crisp c-partitions of X. So, there are four kinds of label vectors, but fuzzy and probabilistic label vectors are mathematically identical, having entries between zero and one that sum to one over each column. These two types of labels, however, are philosophically, conceptually, and computationally different. The reason these matrices are called partitions follows from the interpretation of their entries. If U is crisp or fuzzy, u_{ik} is taken as the membership of $\mathbf{x}_k$ in the ith partitioning fuzzy subset (cluster) of X. If U in M_{fcn} is probabilistic, u_{ik} is usually the (posterior) probability $p(i|\,\mathbf{x}_k)$ that, given $\mathbf{x}_k$, it came from class i. And if U in M_{pcn} is possibilistic, it has entries between zero and one that do not necessarily sum to one over every column. In this last case u_{ik} is taken as the possibility that $\mathbf{x}_k$ belongs to class i. Observe that $M_{hcn} \subset M_{fcn} \subset M_{pcn}$.

An alternative characterization of any U in M_{hcn} is in terms of the c crisp subsets that are defined by the rows of U. Specifically, we may write $X = X_1 \cup \cdots X_i \cup \cdots X_c$, where $X_i \cap X_j = \emptyset$ whenever $i \neq j$. The ith row of U contains a one at each column k where $\mathbf{x}_k$ is in class i and $\sum_{k=1}^{n} u_{ik} = n_i = |X_i|$.

When there is no U in M_{pcn} associated with the data set X, we call it unlabeled data. In this case there are three questions about X:

Q1) Does X have cluster substructure at any value of c, $1 < c < n$?

Q2) If X has substructure, how can we find the clusters?

Q3) Once clusters are found, how can we validate them?

Q1) is called assessment of clustering tendency, and we do not pursue this problem here; see Jain and Dubes [1] or Everitt [2] for formal and informal treatments.

Q2) is called cluster analysis. There are many models and algorithms for clustering based on crisp [3], fuzzy [4], probabilistic [5], and possibilistic methods [6].

Q3) is called cluster validity; once $U(X)$ is found, do we believe it? Better yet, can we use it? Is there a better one we did not find? and so on. Just as tendency assessment depends on how clusters are defined, validation depends on what we mean by a good partition.

While we have specified Q1)–Q3) as if they were straightforward questions, they are vague. For instance, what is meant by cluster structure? Different mathematical properties can be used to define terms like this; they usually lead to rather conflicting ideas about what we think data sets contain. Q1)–Q3) are summarized in Fig. 1, which is a road map of what you can do to and with unlabeled data. It does not include all the special cases and does not tell you what U can be used for once you find it.

In the sequel we concentrate on constrained fuzzy c-partitions of X. The question: which $U \in M_{fcn}$ best explains and represents the (unknown) structure in X? $c = 1$ is represented uniquely by the hard one-partition

$$\mathbf{1}_n = \underbrace{[1 \quad 1 \quad \cdots \quad 1]}_{n\, times}$$

which asserts that all n objects belong to a single cluster. At the other extreme, for $U \in M_{fcn}$, $c = n$ is represented uniquely by $U = I_n$, the $n \times n$ identity matrix, up to a permutation

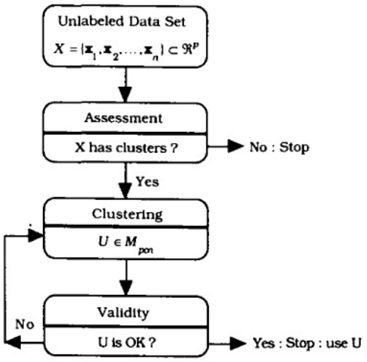

Fig. 1. Processing unlabeled data (exploratory data analysis).

of columns. In this case each object is in its own singleton cluster. Choosing $c = 1$ or $c = n$ rejects the hypothesis that X contains clusters.

Fuzzy clustering algorithms are formally represented as functions $\mathbf{C}: \Re^{p\times n} \mapsto M_{fcn}$. Let $\mathcal{P} = \{U_i: 1 \leq i \leq N\}$ denote N different partitions of a fixed data set X that may arise as a result of clustering X with an algorithm $\mathbf{C}_i$ at various values of its parameters, or more generally, clustering X over different algorithms $\{\mathbf{C}_i\}$, each with its parameters. Each of the U's in $\mathcal{P}$ is a realization of

$$U = \mathbf{C}_i[X; (\underbrace{p_{i1}}_{c}, p_{i2}, \cdots, p_{ik_i})] \quad \text{for some } i \tag{2}$$

where $\{p_{ij}\}$ are the k_i parameters of algorithm $\mathbf{C}_i$. The handful of partitions that you can feasibly generate from an unlabeled data set is a function of the algorithms $\{\mathbf{C}_i\}$ you choose to use, each of which is itself a function of its k_i parameters. The only common denominator of the algorithms $\{\mathbf{C}_i\}$ is the parameter c, the number of clusters to choose; that is why it is explicitly shown in (2). Moreover, for X fixed, c is the most important parameter, in the sense that other parameters of any $\mathbf{C}_i$ really have what we might call second-order effects on U compared to the effect of changing the number of clusters sought in the data. Thus, the most effective strategy for clustering is to first decide what seems to be the most reasonable estimate of the correct number of clusters by choosing one C_i and fixing all of its parameters except c.[1] This results in the problem most often called cluster validity: given

$$\mathcal{P} = \{U_i(c) \in M_{fcn}: U_i(c) = \underbrace{\mathbf{C}_i}_{fixed}[X; (c, \underbrace{p_{i2}, \cdots, p_{ik_i}}_{fixed})]; \\ c = 2, 3, \cdots c_{\max}\} \tag{3}$$

find the best value for c. (There is little guidance in the literature about $c_{\max}$. A rule of thumb that many investigators use is $c_{\max} \leq \sqrt{n}$.)[2]

If C_i is guided by an objective function to the partitions in $\mathcal{P}$, at first glance it seems like values of the objective function should suffice to choose the best one. It is well known, however, that even the global extremum of many objective functions (such as J_1 for hard c-means) can lead to very unrealistic partitions of X (see [4, p. 97] for an example of this behavior). Moreover, some of the intuitively desirable properties that we may want a partition to have cannot be captured by a functional that is easily optimized. These are arguably the two most compelling reasons for introducing cluster validity functionals, $v^*: M_{fcn} \mapsto \Re$ which can be used to rank the validity of various partitions of X.

If $U_i \in M_{hcn}$ is hard, it defines real subsets in X, and various validity criteria such as cluster volume and separability can be measured in an attempt to rank U_i. Measures of validity on M_{hcn} are called direct measures of cluster validity because they assess clusters by examining crisp subsets of the data. Indexes computed on partitions that are not crisp are called indirect validity measures. As written, v^* is suitable for examining partitions generated by any clustering algorithm. In the restricted case shown at (3), one partition will be considered at each value of c, so we may write $v(c) = v^*[U_i(c)]$. In this notation fuzzy validity functionals $v = v^* \circ U_i: \{2, 3, \cdots, c_{\max}\} \mapsto \Re$ depend only on c.

We will further specialize (3) in this paper by choosing $\mathbf{C}_i$ to be any algorithm C_{FCM} that optimizes the FCM (fuzzy c-means) model defined in Section II. The FCM objective function depends on a parameter m, $1 < m < \infty$ called the weighting exponent of the model. Furthermore, J_m is a function of not only U in M_{fcn}, but also an (unknown) vector $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c)$, $\mathbf{v}_i \in \Re^p$, of cluster centers for the fuzzy clusters in X. These are used to specialize (3) as follows:

$$\begin{aligned} \mathcal{P}_{\mathcal{FCM}} &= \{[U_i(c), \mathbf{V}_i(c)] \\ &= \mathbf{C}_{\text{FCM}}[X; (c, m, \underbrace{p_{i3}, \cdots, P_{ik}}_{fixed})]; \\ &\qquad c = 2, 3, \cdots c_{\max}; \quad m \in (1, \infty)\}. \end{aligned} \tag{4}$$

Now we can state the main objective of our paper: once fuzzy clusters $U \in M_{fcn}$ and cluster centers $\mathbf{V}$ in $\Re^{cp}$ are found using

$$\mathbf{C}_{\text{FCM}}[X; (c, m, \underbrace{p_{i3}, \cdots, p_{i7}}_{fixed})]$$

how can we validate them? We will study five methods for doing this. Section II gives a thumbnail sketch of FCM. Section III describes the validity functionals we study and gives an analysis of their behavior for the limiting cases with respect to c. Section IV analyzes the limiting behavior of the validity indexes as m approaches its bounds. Section V contains numerical examples that illustrate the use of and some problems with validation indexes. And Section VI presents our conclusions.

II. The Fuzzy c-Means Clustering Model

The most widely used objective function model for fuzzy clustering in X is the weighted within groups sum of squared errors objective function J_m, which is used to define the

[1] The choice of a particular C_i is guided, whenever possible, by your expectations about possible structural geometries in X.

[2] Strict adherence to this rule is not recommended, however. In most situations there will be some practical limit for $c_{\max}$ that is known to the user.

TABLE I

The Fuzzy c-Means Algorithm (FCM-AO)	
Store	Unlabeled Object Data $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \Re^p$
Pick	☛ $1 < c < n$ ☛ $m > 1$ ☛ T = iteration limit ☛ $0 < \varepsilon$= termination criterion ☛ Norm for J_m : $\langle \mathbf{x}, \mathbf{x} \rangle_A = \|\mathbf{x}\|_A^2 = \mathbf{x}^T A \mathbf{x}$ ☛ Norm for $E_t = \|\mathbf{V}_t - \mathbf{V}_{t-1}\|_{err}$
Guess	$\mathbf{V}_0 = (\mathbf{v}_{1,0}, \mathbf{v}_{2,0}, \ldots, \mathbf{v}_{c,0}) \in \Re^{cp}$
Iterate	For t = 1 to T : Calculate U_t with $\mathbf{V}_{t-1}$ and (6a) Calculate $\mathbf{V}_t$ with U_t and (6b) If $E_t = \|\mathbf{V}_t - \mathbf{V}_{t-1}\|_{err} \le \varepsilon$, Stop and put $(U_f, \mathbf{V}_f) = (U_t, \mathbf{V}_t)$; Else Next t
Use	Prototypes $\mathbf{V}_f$ and/or Fuzzy Labels U_f

constrained optimization problem

$$\underbrace{\min}_{(U,\mathbf{V})} \left\{ J_m(U, \mathbf{V}; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \right\} \quad (5)$$

where $U \in M_{fcn}$, $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c)$ is a vector of (unknown) cluster centers (weights or prototypes), $\mathbf{v}_i \in \Re^p$ for $1 \le i \le c$ and $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}$ is any inner product norm. Optimal partitions U^* of X are taken from pairs $(U^*, \mathbf{V}^*)$ that are local minimizers of J_m. Approximate optimization of J_m by the FCM algorithm is based on iteration through the following necessary conditions for its local extrema.

Fuzzy c-Means Theorem [4]: If $D_{ikA} = \|\mathbf{x}_k - \mathbf{v}_i\|_A > 0$ for all i and k, then $(U, \mathbf{V}) \in M_{fcn} \times \Re^{cp}$ may minimize J_m only if, when m>1

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{D_{ikA}}{D_{jkA}} \right)^{2/(m-1)} \right]^{-1}, \quad 1 \le i \le c; \quad 1 \le k \le n \quad (6a)$$

and

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})^m}, \quad 1 \le i \le c. \quad (6b)$$

Singularity in FCM occurs when one or more of the distances $\|\mathbf{x}_k - \mathbf{v}_i\|_A = 0$ at any iterate. In this case (rare in practice), (6a) cannot be calculated. When this happens, assign zeros to each nonsingular class and distribute memberships to the singular classes arbitrarily subject to the constraints in (1b).

Perhaps the most popular algorithm for approximating solutions of (6) is Picard iteration through (6a) and (6b). This is the algorithm we use to generate approximate solutions of the FCM problem at (5) found in Table I.

This type of iteration is often called alternating optimization (AO) as it simply loops through one cycle of estimates for $\mathbf{V}_{t-1} \rightarrow U_t \rightarrow \mathbf{V}_t$ and then checks $\|\mathbf{V}_t - \mathbf{V}_{t-1}\|_{err} \le \varepsilon$. Equivalently, the entire procedure can be shifted one-half cycle, so that initialization is done on U_0, and the iterates become $U_{t-1} \rightarrow \mathbf{V}_t \rightarrow U_t$, with the alternate termination criterion $\|U_t - U_{t-1}\|_{err} \le \varepsilon$. The literature contains both specifications; the convergence theory is the same in either case. There are some obvious advantages to the form given here in terms of speed and storage. The alternate form that terminates on U's is more stringent, since many more parameters must become close before termination is achieved. It can happen that different results ensue by using the same ε with both forms. The parameter list for FCM-AO is $\{c, m, T, \varepsilon, \|*\|_A, \|*\|_{err}, \mathbf{V}_0\}$. In this study we fix $T = 100$, $\varepsilon = 0.00001$, $\|*\|_A$ is the Euclidean norm, $\|*\|_{err}$ is the one-norm on $\Re^{cp}$, and $\mathbf{V}_0 = c$ randomly chosen distinct points in X.

Conditions (6) are first-order necessary conditions for local extrema of J_m. In principle then, any algorithm $\mathbf{C}_{\text{FCM}}$ used to solve (5) should generate candidates that satisfy (6). For example, you might try to optimize J_m with dynamic programming or perhaps a genetic algorithm; candidate solutions must still satisfy (6). This is an extremely important point for our study because equations (6) are the basis for our limit analysis of the validity functionals studied in Section IV. So, although we will use FCM-AO to generate $(U, \mathbf{V})$ pairs for our numerical examples, what can be learned about the behavior of any v^{FCM} as a function of c and m is independent of the method used to find extrema of J_m.

Some limiting properties of (6) that are important for this study are given in [7]

$$\lim_{m \xrightarrow{+} 1} \left\{ \left[\sum_{j=1}^{c} \left(\frac{D_{ikA}}{D_{jkA}} \right)^{2/(m-1)} \right]^{-1} \right\} = \begin{Bmatrix} 1; & D_{ikA} < D_{jkA} \quad \forall j \ne i \\ 0; & \text{otherwise} \end{Bmatrix}, \quad 1 \le i \le c; \quad 1 \le k \le n. \quad (7a)$$

Using this result, we take the same limit in (6b), obtaining

$$\lim_{m \xrightarrow{+} 1} \left\{ \left[\mathbf{v}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})^m} \right] \right\} = \frac{\sum_{\mathbf{x}_k \in X_i} \mathbf{x}_k}{n_i}, \quad 1 \le i \le c \quad (7b)$$

where $X = X_1 \cup \cdots X_i \cup \cdots X_c$ is the hard c-partition of X defined by the right side of (7a) with $\sum_{k=1}^{n} u_{ik} = n_i = |X_i|$. If we use these results in (5), we have

$$\lim_{m \xrightarrow{+} 1} \left\{ \underbrace{\min}_{(U,\mathbf{V})} \left[J_m(U, \mathbf{V}; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \right] \right\} = \underbrace{\min}_{(U,\mathbf{V})} \left[J_1(U, \mathbf{V}; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \right]. \quad (8)$$

$J_1(U, \mathbf{V}; X)$ is the classical within-groups sum of squared errors objective function. Equation (8) is the hard c-means (HCM) model. Moreover, the right sides of (7a) and (7b) are the necessary conditions for local extrema of J_1. It will be important in Section IV to know that $J_1(U, \mathbf{V}; X)$ can be written as the trace of the within cluster scatter matrix of X when partitioned by U, so we give the definitions here. Let

$X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} = X_1 \cup \cdots X_i \cup \cdots X_c$ be any hard c-partition of X with sample mean vectors $\mathbf{V} = \{\mathbf{v}_i = \sum_{\mathbf{x}_k \in X_i} \mathbf{x}_k/n_i\}$ and grand mean $\overline{\mathbf{v}} = \sum_{\mathbf{x} \in X} \mathbf{x}/n$. For any pair $(U, \mathbf{V}) \in M_{hcn} \times \Re^{cp}$, we define

$$\begin{aligned} S_i &= \left[\sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T\right] \\ &= \text{scatter matrix for } X_i; \end{aligned} \tag{9a}$$

$$\begin{aligned} S_W &= \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \left[\sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T\right] \\ &= \text{within cluster scatter matrix.} \end{aligned} \tag{9b}$$

$$\begin{aligned} S_B &= \sum_{i=1} n_i(\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{v}_i - \overline{\mathbf{v}})^T \\ &= \text{between cluster scatter matrix} \end{aligned} \tag{9c}$$

and

$$\begin{aligned} S_T &= S_W + S_B = \sum_{k=1}^{n} (\mathbf{x}_k - \overline{\mathbf{v}})(\mathbf{x}_k - \overline{\mathbf{v}})^T \\ &= \text{total scatter matirix of } X. \end{aligned} \tag{9d}$$

For any $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}, \operatorname{tr}[\sum_{k=1}^{n} (\mathbf{x}_k)(\mathbf{x}_k)^T] = \sum_{k=1}^{n} ||\mathbf{x}_k||^2$. Using this with (9b)–(9d) yields

$$\begin{aligned} \operatorname{tr}(S_w) &= \operatorname{tr}\left\{\sum_{i=1}^{c} \left[\sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T\right]\right\} \\ &= \sum_{i=1}^{c} \left(\sum_{\mathbf{x}_k \in X_i} ||\mathbf{x}_k - \mathbf{v}_i||^2\right) \\ &= J_1(U, \mathbf{V}; X); \end{aligned} \tag{10}$$

$$\begin{aligned} \operatorname{tr}(S_B) &= \operatorname{tr}\left[\sum_{i=1}^{c} n_i(\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{v}_i - \overline{\mathbf{v}})^T\right] \\ &= \sum_{i=1}^{c} n_i||\mathbf{v}_i - \overline{\mathbf{v}}||^2; \end{aligned} \tag{11}$$

$$\begin{aligned} \operatorname{tr}(S_T) &= \operatorname{tr}\left[\sum_{k=1}^{n} (\mathbf{x}_k - \overline{\mathbf{v}})(\mathbf{x}_k - \overline{\mathbf{v}})^T\right] \\ &= \sum_{k=1}^{n} ||\mathbf{x}_k - \overline{\mathbf{v}}||^2 \\ &= C_X. \end{aligned} \tag{12}$$

The total scatter matrix S_T at (9d) is a function of X alone, so its trace at (12) is the constant C_X which depends only on X. Specifically, $\operatorname{tr}(S_T)$ is not dependent on $(U, \mathbf{V})$. Consequently, for a fixed data set $\operatorname{tr}(S_T) = \operatorname{tr}(S_W) + \operatorname{tr}(S_B) = J_1(U, \mathbf{V}; X) + \operatorname{tr}(S_B) = C_X$, so when we minimize J_1, we simultaneously maximize $\operatorname{tr}(S_B)$ which is a measure of the between cluster scatter of any $(U, \mathbf{V}) \in M_{hcn} \times \Re^{cp}$.

Lastly, we need the limits of (6a) and (6b) as m approaches infinity [7]

$$\lim_{m \to \infty} \left\{\left[\sum_{j=1}^{c} \left(\frac{D_{ikA}}{D_{jkA}}\right)^{2/(m-1)}\right]^{-1}\right\} = \frac{1}{c}, \quad 1 \le i \le c; \quad 1 \le k \le n; \tag{13a}$$

$$\lim_{m \to \infty} \left\{\left[\mathbf{v}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})^m}\right]\right\} = \frac{\sum_{k=1}^{n} \mathbf{x}_k}{n} = \overline{\mathbf{v}}, \quad 1 \le i \le c \tag{13b}$$

where $\overline{\mathbf{v}}$ is again the grand mean of X.

III. Cluster Validity for U in M_{fcn}

Validity methods associated with, but not specifically designed for, the FCM model began with Bezdek's partition coefficient v_{PC} [8] and partition entropy v_{PE} [9] of any U in M_{fcn}

$$v_{PC}(U) = \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2}{n} \tag{14}$$

and

$$v_{PE}(U) = -\frac{1}{n}\left\{\sum_{k=1}^{n} \sum_{i=1}^{c} [u_{ik} \log_a (u_{ik})]\right\}. \tag{15}$$

In (15) $a \in (1, \infty)$ is the logarithmic base. Properties of these two indexes as functions of U and c were studied in [8]–[10]. We repeat the main results

$$\begin{aligned} v_{PC}(U) = 1 \Leftrightarrow v_{PE}(U) &= 0 \Leftrightarrow U \in M_{hcn} \\ &\quad \text{is a } hard\, c\text{-partition of } X; \end{aligned} \tag{16}$$

$$\begin{aligned} v_{PC}(U) = \frac{1}{c} \Leftrightarrow v_{PE}(U) &= \log_a(c) \Leftrightarrow U \\ &= \left[\frac{1}{c}\right] \doteq \overline{U}. \end{aligned} \tag{17}$$

Equation (16) shows that v_{PC} takes its maximum (and v_{PE} takes its minimum) on every hard c-partition. And v_{PC} takes its unique minimum (and v_{PE} takes its unique maximum) at the centroid $U = [1/c] \doteq \overline{U}$ of M_{fcn}. $\overline{U}$ is the "fuzziest" partition you can get, since it assigns every point in X to all c classes with equal membership values $1/c$. These two indexes essentially measure the distance U is from being crisp (i.e., they measure the fuzziness in U). Normalizations of both indexes based on nonstatistical [10] and statistical [11] criteria help reduce their tendency toward being monotonic with c.

In the context of validation, it is clear that when an algorithm produces a partition U that is close to $\overline{U}$, that algorithm is not finding very good cluster substructure in X. This may be the fault of the algorithm, or the data may lack structure. Consequently, the unique minimum v_{PC} (or maximum v_PE) are very helpful in deciding when the structure is not being found. It is less clear that when U approaches M_{hcn}, cluster substructure has been found. Since $v_{PC} = 1$ $(v_{PE} = 0)$ for every U in M_{hcn}, it is incorrect to assert that just because v_{PC} is near one (or v_{PE} is near zero) that U is a good clustering of X. Empirical studies vary: some have shown that maximizing v_{PC} (or minimizing v_{PE}) over $\mathcal{P}$ at (3) or $\mathcal{P}_{\mathcal{FCM}}$ at (4) often (but not always) leads to a good interpretation of the data [8]–[11]; others have shown that different indexes are sometimes more effective [12]–[14]. This simply confirms what we already know: no matter how good your index is, there is a data set out there waiting to trick it (and you).

υ_{PC} and υ_{PE} are examples of a class of validity functionals that are functions of U in M_{fcn} alone. A strong criticism of indexes like these is that, while $U = \mathbf{C}_i(X)$ is a function of the data set X, functionals such as υ_{PC} and υ_{PE} are only implicitly functions of X. In other words, they do not use the data itself. Moreover, many algorithms generate collateral information or parameters—e.g., the cluster centers $\mathbf{V}$ from the FCM model—that may be useful for validation.

Gunderson's separation coefficient [15] was the first validity index that explicitly used the three components $(U, \mathbf{V}; X)$ where $U \in M_{fcn}$ and $\mathbf{V}$ is a vector of c prototypes that are associated with the clusters in U. More recent indexes in this class are the Xie–Beni [16] indexes υ_{XB}, $\upsilon_{XB,m}$ and the Fukayama–Sugeno [17] index υ_{FS}.

Let $U \in M_{fcn}$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c)$ be a vector of distinct points $\mathbf{v}_i \in \Re^p$ for $1 \le i \le c$ (for us they will usually be cluster centers), and let $||\mathbf{x}||^2 = \mathbf{x}^T\mathbf{x}$ be the Euclidean norm. The Xie–Beni index υ_{XB} is defined as

$$\upsilon_{XB}(U, \mathbf{V}; X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2||\mathbf{x}_k - \mathbf{v}_i||^2}{n\,(\underbrace{\min}_{i\neq j}\{\mathbf{v}_i - \mathbf{v}_j||^2\})} = \left[\frac{\left(\frac{\sigma}{n}\right)}{\operatorname{sep}(\mathbf{V})}\right]. \tag{18}$$

Xie and Beni explained this index by writing it as the ratio of the total variation σ of $(U, \mathbf{V})$ and the separation $\operatorname{sep}(\mathbf{V})$ of the vectors $\mathbf{V}$

$$\sigma(U, \mathbf{V}; X) = \sum_{i=1}^{c}\left(\sum_{k=1}^{n} u_{ik}^2||\mathbf{x}_k - \mathbf{v}_i||^2\right); \tag{19}$$

$$\operatorname{sep}(\mathbf{V}) = \underbrace{\min}_{i\neq j}\{||\mathbf{v}_i - \mathbf{v}_j||^2\}. \tag{20}$$

Note that if $(U, \mathbf{V})$ is an extrema of J_2, then $\sigma = J_2$. A good $(U, \mathbf{V})$ pair should produce a small value of σ because u_{ik} is expected to be high when $||\mathbf{x}_k - \mathbf{v}_i||$ is low. And well-separated $\mathbf{v}_i$'s will produce a high value of $\operatorname{sep}(\mathbf{V})$. So, when $\upsilon_{XB}(U_1, \mathbf{V}_1; X) < \upsilon_{XB}(U_2, \mathbf{V}_2; X)$ for either of these reasons (or both), U_1 is presumably a better partition of X than U_2. Consequently, the minimum of υ_{XB} over $\mathcal{P}$ at (3) or $\mathcal{P}_{\mathcal{FCM}}$ at (4) is taken as the most desirable partition of X.

Xie and Beni state that υ_{XB} decreases monotonically when c is close to n. Assume there is only one candidate pair $(U_c, \mathbf{V}_c)$ at each $c = 2, 3, \cdots, n-1$. To avoid monotonicity, they recommend plotting $\upsilon_{XB}(U_c, \mathbf{V}_c; X)$ as a function of c and then selecting the starting point of the monotonic epoch as the maximum c ($c_{\max}$) to be considered. Then the optimum value of c is obtained by minimizing $\upsilon_{XB}(U_c, \mathbf{V}_c; X)$ over $c = 2, 3, \cdots, c_{\max}$.

When $(U, \mathbf{V})$ pairs optimize a criterion function (J) which is very different from J_2, Xie and Beni recommend modifying σ to be compatible with J. In particular, they recommend replacing σ with J_m when $(U, \mathbf{V})$ pairs optimize the FCM model for $m \neq 2$. We will call this the extended FCM Xie–Beni index $\upsilon_{XB,m}^{\mathrm{FCM}}$

$$\upsilon_{XB,m}^{\mathrm{FCM}}(U, \mathbf{V}; X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m||\mathbf{x}_k - \mathbf{v}_i||^2}{n\,(\underbrace{\min}_{i\neq j}\{||\mathbf{v}_i - \mathbf{v}_j||^2\})} = \frac{J_m(U, \mathbf{V}; X)}{n\,(\underbrace{\min}_{i\neq j}\{||\mathbf{v}_i - \mathbf{v}_j||^2\})}. \tag{21}$$

Another functional that combines the three components $(U, \mathbf{V}; X)$ is the validity function $\upsilon_{FS,m}$ of Fukuyama and Sugeno [17]

$$\begin{aligned}\upsilon_{FS,m}(U, \mathbf{V}{:}\,X) &= \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \cdot \left(||\mathbf{x}_k - \mathbf{v}_i||_A^2 - ||\mathbf{v}_i - \overline{\mathbf{v}}||_A^2\right)\\ &= J_m(U, \mathbf{V}{:}\,X) \underbrace{-\sum_{i=1}^{c}\left\{\left[\sum_{k=1}^{n}(u_{ik})^m\right]||\mathbf{v}_i - \overline{\mathbf{v}}||_A^2\right\}}_{K_m(U, \mathbf{V}; X)},\\ &= J_m(U, \mathbf{V}{:}\,X) - K_m(U, \mathbf{V}{:}\,X)\end{aligned} \tag{22}$$

where $1 < m < \infty$. While $\upsilon_{FS,m}$ was not explicitly designed for searching only $\mathcal{P}_{\mathcal{FCM}}$, it seems likely that this was the motivation for its definition. The first term in (22) is J_m, which combines the fuzziness in U with the geometrical compactness of the representation of X via the c prototypes $\mathbf{V}$. The second term, K_m in (22), combines the fuzziness in each row of U with the distance from the ith prototype to the grand mean of the data. Fukuyama and Sugeno proposed that minima of $\upsilon_{FS,m}$ over $\mathcal{P}_{\mathcal{FCM}}$ at (4) pointed to good clusterings of X (good in the sense of minimizing $\upsilon_{FS,m}$, of course). $\upsilon_{FS,m}$ has been used to select the number of rules for a fuzzy controller [18]. When $(U, \mathbf{V})$ is a pair that optimizes the FCM model and $J_m(U, \mathbf{V}_1; X) < J_m(U_2, \mathbf{V}_2; X)$, $U_1$is likely to be a more desirable partition of X than U_2 (but, like J_1, there are examples where this is false [8]). The interpretation of K_m is not so obvious, but taking its limits with m will enable us to get an idea of the meaning of this term.

IV. Limiting Behavior of the Indexes on FCM Pairs

In this section $(U, \mathbf{V})$ pairs are assumed to be optimal for the FCM model at (5), so that they satisfy necessary conditions (6) and limit conditions (7) and (13). We will indicate this by the notation $\upsilon^{\mathrm{FCM}}(U, \mathbf{V}; X)$. Although indexes such as $\upsilon_{XB,m}$ and $\upsilon_{FS,m}$ do incorporate collateral information about cluster substructure that resides in $\mathbf{V}$, when $(U, \mathbf{V})$ pairs minimize J_m, care must be taken to account for their limiting behavior as a function of m. We will show that $\upsilon_{XB,m}^{\mathrm{FCM}}$ and $\upsilon_{FS,m}^{\mathrm{FCM}}$ can be strongly influenced by m. Paradoxically, it is precisely because these two indexes use (FCM) centroids that they are sensitive to m.

Mindful of (7) and (13), we now take limits of the validity functionals as m approaches one from above or infinity. The results for the partition coefficient and entropy hold no surprises.

Partition Coefficient, $m \xrightarrow{+} 1$ and $m \to \infty$:

$$\lim_{m \xrightarrow{+} 1} \{v_{PC}^{\text{FCM}}(U, \mathbf{V}; X)\} = 1;$$
$$\lim_{m \to \infty} \{v_{PC}^{\text{FCM}}(U, \mathbf{V}; X)\} = 1/c. \tag{23}$$

Partition Entropy, $m \xrightarrow{+} 1$ and $m \to \infty$:

$$\lim_{m \xrightarrow{+} 1} \{v_{PE}^{\text{FCM}}(U, \mathbf{V}; X)\} = 0;$$
$$\lim_{m \to \infty} \{v_{PE}^{\text{FCM}}(U, \mathbf{V}; X)\} = \log_a c. \tag{24}$$

These two indexes are independent of $\mathbf{V}$, so their dependency on m seems transparent. Our examples, however, will show what these limits suggest. First, for values of m very close to one both indexes lose their ability to discriminate between various values of c. This happens because the first limits in (23) and (24) take the same value on all crisp U's for every c. At the other extreme, when m becomes large, they will both select $c = 2$ because of the second limits in (23) and (24). That is, for example, the partition coefficient will maximize at one-half because as m approaches infinity, the second limit at (23) yields

$$\frac{1}{2} = \underbrace{\max}_{2 \le c} \left\{\frac{1}{c}\right\}.$$

Xie–Beni Index and the Extended FCM Xie–Beni Index, $m \xrightarrow{+} 1$:

$$\lim_{m \xrightarrow{+} 1} \{v_{XB}^{\text{FCM}}(U, \mathbf{V}; X)\} = \lim_{m \xrightarrow{+} 1} \{v_{XB,m}^{\text{FCM}}(U, \mathbf{V}; X)\} = \frac{J_1(U, \mathbf{V}; X)}{n\,(\underbrace{\min}_{i \neq j} \{\|\mathbf{v}_i - \mathbf{v}_j\|^2\})}. \tag{25}$$

Since the denominator in (25) varies with $\mathbf{V}$ alone, it is very possible that this index will validate different $(U,\mathbf{V})$ pairs than $J_1(U, \mathbf{V})$ does. On the other hand, when we take the limit of the Fukuyama–Sugeno Index, this is not the case.

The Fukuyama–Sugeno Index, $m \xrightarrow{+} 1$:

$$\begin{aligned}
&\lim_{m \xrightarrow{+} 1} \{v_{FS,m}^{\text{FCM}}(U, \mathbf{V}; X)\} \\
&= \lim_{m \xrightarrow{+} 1} \{J_m(U, \mathbf{V}{:}\, X) - K_m(U, \mathbf{V}{:}\, X)\} \\
&= \lim_{m \xrightarrow{+} 1} \{J_m(U, \mathbf{V}{:}\, X)\} - \lim_{m \xrightarrow{+} 1} \{K_m(U, \mathbf{V}{:}\, X)\} \\
&= \lim_{m \xrightarrow{+} 1} \left\{\sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|^2\right\} \\
&\quad - \lim_{m \xrightarrow{+} 1} \left\{\sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|\mathbf{v}_i - \overline{\mathbf{v}}\|^2\right\}.
\end{aligned}$$

Now as $m \xrightarrow{+} 1$, every $u_{ik} \to 0$ or 1 as in (7a). Consequently

$$\lim_{m \xrightarrow{+} 1} \{v_{FS,m}^{\text{FCM}}(U, \mathbf{V}; X)\} = \sum_{i=1}^{c} \left(\sum_{\mathbf{x}_k \in X_i} \|\mathbf{x}_k - \mathbf{v}_i\|^2\right) - \sum_{i=1}^{c} \left(\sum_{\mathbf{x}_k \in X_i} \|\mathbf{v}_i - \overline{\mathbf{v}}\|^2\right).$$

Applying (10) to the first term and (11) to the second term now gives

$$\begin{aligned}
&\lim_{m \xrightarrow{+} 1} \{v_{FS,m}^{\text{FCM}}(U, \mathbf{V}; X)\} \\
&= J_1(U, \mathbf{V}{:}\, X) - \operatorname{tr}\left[\sum_{i=1}^{c} n_i (\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{v}_i - \overline{\mathbf{v}})^T\right] \\
&= \operatorname{tr}(S_W) - \operatorname{tr}(S_B) \\
&= \operatorname{tr}(S_W) - [\operatorname{tr}(S_T) - \operatorname{tr}(S_W)] \\
&= 2\operatorname{tr}(S_W) - \operatorname{tr}(S_T) \\
&= 2J_1(U, \mathbf{V}{:}\, X) - C_X.
\end{aligned} \tag{26}$$

In (26) S_W and S_B are the within cluster and between clusters scatter matrices, respectively, of the limiting hard partition of X shown in (7a). Equation (26) invites two comments for m close to one:

i) J_m is in some sense the fuzzy analog of the within cluster scatter, and K_m is the fuzzy analog of the between cluster scatter.

ii) $v_{FS,m}^{\text{FCM}}$ behaves very much like J_1, which does not necessarily serve us well as a validity functional. In other words, K_m—a measure of the between cluster scatter—has a very negligible effect on the evaluation of $(U, \mathbf{V})$ for very small values of m.

Next, using the results at (13), we take limits as m approaches infinity.

Xie–Beni Index, $m \to \infty$: Recall from (13b) that $\lim_{m \to \infty} \{\mathbf{v}_i\} = \overline{\mathbf{v}}$. For the numerator of the Xie–Beni index we have

$$\begin{aligned}
&\underbrace{\lim}_{m \to \infty} \left\{\frac{\sigma(U, \mathbf{V}; X)}{n}\right\} \\
&= \frac{1}{n} \underbrace{\lim}_{m \to \infty} \left\{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2 \|\mathbf{x}_k - \mathbf{v}_i\|^2\right\} \\
&= \left(\frac{1}{nc^2}\right) \sum_{i=1}^{c} \sum_{k=1}^{n} \|\mathbf{x}_k - \overline{\mathbf{v}}\|^2 \\
&= \left(\frac{c}{nc^2}\right) \sum_{k=1}^{n} \|\mathbf{x}_k - \overline{\mathbf{v}}\|^2 \\
&= \left(\frac{1}{nc}\right) \operatorname{tr}(S_T) \\
&= \frac{C_X}{nc}.
\end{aligned}$$

For the denominator in (18), however, we have

$$\begin{aligned}\underbrace{\lim}_{m\to\infty}\{\operatorname{sep}(\mathbf{V})\} &= \underbrace{\lim}_{m\to\infty}(\underbrace{\min}_{i\neq j}\{\|\mathbf{v}_i-\mathbf{v}_j\|^2\})\\ &=(\underbrace{\min}_{i\neq j}\{\|\overline{\mathbf{v}}-\overline{\mathbf{v}}\|^2\})\\ &=0.\end{aligned}$$

Consequently

$$\lim_{m\to\infty}\{v_{XB}^{\mathrm{FCM}}(U,\mathbf{V};X)\}=\frac{C_X/nc}{0}=\infty. \tag{27}$$

This leads, and our examples will confirm, to numerical instability for large values of m.

The Extended FCM Xie–Beni, $m \to \infty$: The extended FCM Xie–Beni index $v_{XB,m}^{\mathrm{FCM}}(U,\mathbf{V};X)=\{J_m\,(U,\mathbf{V};X)/n\,[\operatorname{sep}(\mathbf{V})]\}$ has a different limiting behavior than v_{XB}^{FCM} because the power of u_{ik} in the numerator is m, not fixed at two. The denominator still goes to zero, but with $(u_{ik})^m$ in the numerator, it also goes to zero. To see this we take the limit of the FCM objective function using (13) and the Euclidean norm

$$\begin{aligned}&\underbrace{\lim}_{m\to\infty}\left\{J_m(U,\mathbf{V}:X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m\|\mathbf{x}_k-\mathbf{v}_i\|^2\right\}\\ &=\left(\sum_{k=1}^{n}\|\mathbf{x}_k-\overline{\mathbf{v}}\|^2\right)\lim_{m\to\infty}\left\{\frac{c}{c^m}\right\}\\ &=C_X\left(\lim_{m\to\infty}\left\{\frac{1}{c^{m-1}}\right\}\right)\\ &=0.\end{aligned}$$

This leads to the indeterminate form

$$\lim_{m\to\infty}\{v_{XB,m}^{\mathrm{FCM}}(U,\mathbf{V};X)\}=\frac{0}{0} \tag{28}$$

so the behavior of $v_{XB,m}^{\mathrm{FCM}}(U,\mathbf{V};X)$ for large values of m is not at all predictable.

The Fukuyama–Sugeno Index, $m \to \infty$: Since

$$\begin{aligned}\lim_{m\to\infty}\{K_m(U,\mathbf{V}:X)\}&=\|\overline{\mathbf{v}}-\overline{\mathbf{v}}\|^2\lim_{m\to\infty}\left\{\frac{1}{c^{m-1}}\right\}\\ &=0\cdot 0=0\end{aligned}$$

we have

$$\begin{aligned}&\lim_{m\to\infty}\{v_{FS,m}^{\mathrm{FCM}}(U,\mathbf{V};X)\}\\ &\quad=\lim_{m\to\infty}\{J_m(U,\mathbf{V}:X)-K_m(U,\mathbf{V}:X)\}\\ &\quad=0-(0\cdot 0)=0. \end{aligned} \tag{29}$$

Consequently, this index loses its ability to validate $(U, \mathbf{V})$ pairs from FCM for large m.

To summarize, very low or high values of m may influence any validity index that uses membership values U from FCM. Moreover, indexes that also use $\mathbf{V}$ from FCM may experience additional problems because $\lim_{m\to\infty}\{\mathbf{v}_i\}=\overline{\mathbf{v}}$. And indexes such as $v_{XB,m}^{\mathrm{FCM}}$ or $v_{FS,m}^{\mathrm{FCM}}$ that are explicit functions of m and $(U, \mathbf{V})$ may be very unreliable for small or large values of m. For example, dependency on m makes $v_{FS,m}^{\mathrm{FCM}}$ or $v_{FS,m}^{\mathrm{FCM}}$ unreliable for high values of m. Finally, $v_{FS,m}^{\mathrm{FCM}}$ is also unpredictable for low values of m (very near one). Next we will illustrate the force of these remarks with several numerical examples.

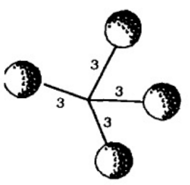

Fig. 2. Schematic representation of Normal-4.

V. Numerical Examples on Cluster Validity

Perhaps you are wondering what a low or high value of m is? After all, the limits in Section IV are hardly applicable in practical situations. Most users of FCM choose values of m in the range (1, 10] (we are aware of one or two studies that have examined m up to about 30, but values of $m < 5$ or so are more usual, and $m = 2$ is by far the most common choice).

To study the indexes, we use two data sets named Normal-4 and IRIS. Normal-4 is a sample of $n = 800$ points consisting of 200 points each from the four components of a mixture of $c = 4$, $p = 4$-variate normals. The population mean vector and covariance matrix for each component of the normal mixture were $\mu_i = 3\mathbf{e}_i$ and $\Sigma_i = I_4$, $i = 1, 2, 3, 4$, $\mathbf{e_i} = (0, \cdots, \underbrace{1}_{i}, \cdots, o)$.

It is helpful to picture the geometric structure of Normal-4, which is (a sample of) 200 points each centered at three units from the origin along each of the four coordinate axes, with unit variance for each sample in all four directions. Fig. 2 shows what this data looks like to the mind's eye if the sampling of each component is very nice. Because the standard deviation of each population component is one, we can only expect about 68.2% of each 200 samples to be within one unit of their mean. Just add another axis in your mind to visualize Normal-4.

IRIS has $n = 150$ points in $p = 4$ dimensions that represent three physical clusters each with 50 points [19]. We say physical because although IRIS contains observations from three different physical classes of flowers, in their numerical representation two of the classes have substantial overlap, while the third is well separated from the other two. Thus, one can argue in favor of both $c = 2$ and $c = 3$ for IRIS.

For each data set we made several runs of FCM for different values of m. As a reminder, in this study all other parameters of FCM were fixed: $T = 100$, $\varepsilon = 0.00001$, $\|*\|_A$ is the Euclidean norm, $\|*\|_{err}$ is the one-norm on $\Re^{cp}$, $\mathbf{V}_0 = c$ randomly chosen distinct points in X. For a particular c and data set the same initial centroids were used for all runs. Experiments have also been done with different initializations not reported here; those results were very similar to the ones given.

Table II displays the values of the five validity indexes for $c = 2$ to 10 for terminal $(U, \mathbf{V})$ FCM pairs of NORMAL-4 for the weighting exponents $m = 1.2$ and $m = 7$. We have highlighted and shaded the optimal value of c chosen by

TABLE II
INDEX VALUES ON NORMAL-4 FOR $c = 2$ TO 10: $m = 1.2$ AND $m = 7$

c	J_m	K_m	$v_{FS,m}^{FCM} = J_m - K_m$	$2J_m - C_X$	v_{XB}^{FCM}	$v_{XB,m}^{FCM}$	v_{PC}^{FCM}	v_{PE}^{FCM}
	***m* = 1.2**							
2	4777	1722	3054	990	0.55	0.60	0.89	0.19
3	1983	3573	- 1590	- 4596	0.16	0.20	0.92	0.15
4	2320	5442	- 3122	- 3923	**0.16**	**0.17**	**0.95**	**0.09**
5	2128	5558	- 3429	- 4307	0.41	0.47	0.92	0.15
6	2106	5649	- 3542	- 4350	0.68	0.80	0.89	0.20
7	2015	5736	- 3720	- 4533	0.70	0.85	0.86	0.25
8	2175	5870	- 3694	- 4213	0.67	0.80	0.87	0.25
9	1971	5929	- 3958	- 4621	0.67	0.85	0.83	0.30
10	1882	6015	**- 4132**	- 4798	0.66	0.83	0.84	0.31
	***m* = 7**							
2	66.91	0.00	66.91	NA	**1.87 E06**	58468	**0.50**	**0.69**
3	7.83	0.00	7.83	NA	2.48 E06	10224	0.33	1.10
4	1.57	0.00	1.57	NA	8.74 E06	8543	0.25	1.39
5	0.44	0.00	0.44	NA	3.34 E06	1069	0.20	1.61
6	0.15	0.00	0.15	NA	3.22 E06	414	0.17	1.79
7	0.06	0.00	0.06	NA	8.86 E06	527	0.14	1.95
8	0.03	0.00	0.03	NA	1.17 E06	357	0.13	2.08
9	0.01	0.00	0.01	NA	9.55 E06	161	0.11	2.20
10	0.01	0.00	**0.01**	NA	9.63 E06	**96**	0.10	2.30

TABLE III
INDEX VALUES ON IRIS FOR $c = 2$ TO 10: $m = 1.2$ AND $m = 7$

c	J_m	K_m	$v_{FS,m}^{FCM} = J_m - K_m$	$2J_m - C_X$	v_{XB}^{FCM}	$v_{XB,m}^{FCM}$	v_{PC}^{FCM}	v_{PE}^{FCM}
	***m* = 1.2**							
2	25.99	529.49	- 503	- 629	**0.01**	**0.01**	**0.99**	**0.02**
3	62.59	600.72	- 538	- 556	0.12	0.13	0.98	0.04
4	46.89	619.93	- 573	- 587	0.18	0.19	0.96	0.06
5	46.18	623.64	- 577	- 589	0.48	0.51	0.95	0.08
6	44.74	624.34	- 579	- 591	0.72	0.77	0.95	0.09
7	45.55	625.49	- 579	- 590	1.12	1.20	0.95	0.09
8	44.58	625.33	- 580	- 592	2.05	2.19	0.95	0.09
9	30.46	637.34	- 606	- 620	1.40	1.50	0.96	0.08
10	25.79	641.37	**- 615**	- 629	1.16	1.27	0.94	0.11
	***m* = 7**							
2	3.03	29.69	**- 26.65**	NA	**0.10**	0.0013	**0.54**	**0.65**
3	0.35	4.93	- 4.58	NA	0.62	0.0016	0.37	1.05
4	0.07	1.17	- 1.10	NA	1.13	0.0006	0.27	1.34
5	0.02	0.22	- 0.20	NA	7.98	0.0015	0.22	1.56
6	0.01	0.05	- 0.04	NA	26326	2.0163	0.18	1.75
7	0.00	0.03	- 0.03	NA	32470	1.0905	0.16	1.90
8	0.00	0.13	- 0.12	NA	808502	14.4105	0.14	2.03
9	0.00	2.05	- 2.05	NA	17599	0.1645	0.13	2.12
10	0.00	7.35	- 7.35	NA	2.54	**0.0000**	0.14	2.19

each index (recall that they are all to be minimized except the partition entropy). Since the desired value is $c = 4$, we see that four of the five indexes point to the correct choice at $m = 1.2$, and none of them indicate $c = 4$ at $m = 7$. Observe that while v_{XB}^{FCM} agrees (incorrectly) with v_{PC}^{FCM} and v_{PE}^{FCM} at $m = 7$, the extended Xie–Beni index $v_{XB,m}^{\mathrm{FCM}}$ points to $c = 10$; this behavior is consistent with the fact that limits at (27) and (28) are different. Moreover, the values of $v_{XB,m}^{\mathrm{FCM}}$ at $m = 7$ show a dramatic decrease with c. From the trend in this column, you might suspect that c is approaching zero much faster than Xie and Beni conjectured that it will, since $n = 800$ here. Another point worth considering is the magnitudes of v_{XB}^{FCM} at $m = 7$, which are very large. This is due to the behavior of the denominator shown at (27); some of the values $||\mathbf{v}_i - \mathbf{v}_j|| \to 0$.

The column headed $(2J_m - C_X)$ in Table II is shown adjacent to $v_{FS,m}^{\mathrm{FCM}}$ to check the limit in (26), which asserts that $v_{FS,m}^{\mathrm{FCM}} \to (2J_m - C_X)$ as $m \to 1$ from above. As is evident, the values are not very close (although they certainly exhibit the same trends). We will study this aspect of the limit process in Table IV. At the other extreme, we may conclude that for this data set and these algorithmic parameters $m = 7$ is too high for good FCM pairs. Notice especially that the Fukuyama–Sugeno index $v_{FS,m}^{\mathrm{FCM}}$ points to $c = 10$ for both values of m! Other $(U, \mathbf{V})$ pairs from FCM using, for example, different initializations, could certainly alter these results. But for this example, and for others like it not reported here, our experience has been that the Fukuyama–Sugeno index $v_{FS,m}^{\mathrm{FCM}}$ is very unstable for low and high values of m.

Table III lists the outputs of the five indexes on $(U, \mathbf{V})$ pairs from FCM applied to the IRIS data for the same two values of m. For $m = 1.2$ all indexes point to $c = 2$ except the Fukuyama–Sugeno index, which again points to $c = 10$. In view of our remarks about the geometric structure of IRIS, we take $c = 2$ as a good choice, so the behavior of the five indexes at $m = 1.2$ is exactly the same for the data sets Normal-4 and IRIS. We have again shown values for $2J_m - C_X$, the limit of $v_{FS,m}^{\mathrm{FCM}}$ from above, and again, we see that they are not very close.

The behavior of the Xie–Beni indexes as functions of c for $m = 7$ is very peculiar. For example, v_{XB}^{FCM} is 0.10 for $c = 2$, grows to 808 502 at $c = 8$, and then plummets back to nearly zero (2.54 is nearly zero relative to the values in this column) at $c = 10$. Looking at the trend of values, one is left with the very correct impression that there is a very strong and unpredictable interaction between c and m; moreover, this behavior is not consistent across different data sets. These remarks should serve as a strong warning about what you can and cannot infer from cluster validation indexes.

Unlike their performance in Table II, not all of the indexes fail for IRIS. Indeed, v_{XB}^{FCM}, v_{PC}^{FCM}, and v_{PE}^{FCM} all secure $c = 2$ at both values of m. This again emphasizes how important it is to remember that the data set X determines the quality of inferences that can be made from validity studies. Normal-4 is a fairly well-structured data set, but all indexes fail for $m = 7$, while three work well for IRIS at $m = 7$. The partition coefficient and entropy exhibit the behavior shown in the second limits at (23) and (24) for IRIS. The partition coefficient, for example, maximizes at 0.54, a little above $1/c = 1/2$. All other values of v_{PC}^{FCM} are similar—just a few hundreths above $1/c$. We should observe that while the partition coefficient and entropy both indicate $c = 2$ (the preferred value) for IRIS at every m in Table III, there is some m above which, because of this limiting behavior, that their success is accidental.

Equation (26) asserts that $v_{FS,m}^{\mathrm{FCM}} \to (2J_m - C_X)$ as $m \to 1$ from above. How close to one is close enough to see this behavior even approximately? Table IV compares values of $v_{FS,m}^{\mathrm{FCM}}$ to $2J_m - C_X$ for $m = 1.2$, 1.01, and 1.005 on $(U, \mathbf{V})$ pairs generated by FCM.

TABLE IV
VALUES OF $v_{FS,m}^{\mathrm{FCM}}$ AND $2J_m - C_X$ ON IRIS: $m = 1.2$, 1.05, AND 1.005

	***m* = 1.2**		***m* = 1.01**		***m* = 1.005**	
c	$v_{FS,m}^{FCM}$	$2J_m - C_X$	$v_{FS,m}^{FCM}$	$2J_m - C_X$	$v_{FS,m}^{FCM}$	$2J_m - C_X$
2	- 503	- 629	- 500	- 624	- 661	- 675
3	- 538	- 556	- 538	- 553	- 940	- 954
4	- 573	- 587	- 576	- 586	- 930	- 932
5	- 577	- 589	- 586	- 591	- 928	- 929
6	- 579	- 591	- 588	- 591	- 931	- 931
7	- 579	- 590	- 589	- 592	**- 946**	- 955
8	- 580	- 592	- 589	- 590	- 938	- 939
9	- 606	- 620	- 615	- 626	- 939	- 939
10	**- 615**	- 629	**- 623**	- 626	- 941	- 939

TABLE V
OPTIMAL VALUES OF c FOR NORMAL-4 AND IRIS CHOSEN BY EACH INDEX: $c = 2$–10

Normal-4 : preferred value = 4					
m	$v_{FS,m}^{FCM}$	v_{XB}^{FCM}	$v_{XB,m}^{FCM}$	v_{PC}^{FCM}	v_{PE}^{FCM}
1.01	10	**4**	**4**	(?)	(?)
1.1	10	**4**	**4**	**4**	**4**
1.2	10	**4**	**4**	**4**	**4**
1.5	**4**	**4**	**4**	**4**	**4**
1.8	**4**	**4**	**4**	**4**	2
2.0	**4**	**4**	**4**	2	2
2.2	**4**	**4**	**4**	2	2
2.5	**4**	**4**	**4**	2	2
2.8	**4**	**4**	**4**	2	2
3.0	10	**4**	**4**	2	2
3.5	10	2	10	2	2
4.0	10	3	10	2	2
7.0	10	2	10	2	2
IRIS: preferred value = 2					
m	$v_{FS,m}^{FCM}$	v_{XB}^{FCM}	$v_{XB,m}^{FCM}$	v_{PC}^{FCM}	v_{PE}^{FCM}
1.01	10	**2**	**2**	(?)	(?)
1.1	10	**2**	**2**	**2**	**2**
1.2	10	**2**	**2**	**2**	**2**
1.5	4	**2**	**2**	**2**	**2**
1.8	3	**2**	**2**	**2**	**2**
2.0	**2**	**2**	**2**	**2**	**2**
2.2	**2**	**2**	**2**	**2**	**2**
2.5	**2**	**2**	**2**	**2**	**2**
2.8	**2**	**2**	**2**	**2**	**2**
3.0	**2**	**2**	**2**	**2**	**2**
3.5	**2**	**2**	**2**	**2**	**2**
4.0	**2**	**2**	**2**	**2**	**2**
7.0	**2**	**2**	10	**2**	**2**

Even at 1.005 there are values of c for which $|v_{FS,m}^{\mathrm{FCM}} - (2J_m - C_X)|$ is not small. For example, at $c = 7$ this difference is nine. So, limit really means limit. Note, however, that $v_{FS,m}^{\mathrm{FCM}}$ selects $c = 7$ for $m = 1.005$, while it selects $c = 10$ at $m = 1.2$ and 1.01. This again illustrates how sensitive this index is to changes in m.

Finally, Table V lists the value of c chosen by each of the five validity indexes for 13 values of m ranging from 1.01–7 for each of our data sets, when c ranged over the integers from 2–10. We have shaded those cells of the table that agree with the preferred value of c for each data set. The values in Table V invite several comments.

i) *The Partition Coefficient and Partition Entropy*: At $m = 1.01$ (and any m closer to one than this) the values of v_{PC}^{FCM} and v_{PE}^{FCM} on every $(U, \mathbf{V})$ pair from $c = 2$ to $c = 10$ were identical for each data set (separately). Why? For values of m very close to one, both indexes lose their ability to discriminate between various values of c because the first limits in (23) and (24) take the same value on all crisp U's for every c. This results in a tie (up to this level of accuracy), so we call these undecidable cases, marked in the table by (?). What this indicates is that for $m = 1.01$, the partitions U from FCM $(U, \mathbf{V})$ pairs are so close to some vertex of M_{hcn} that they are, for all practical purposes, crisp partitions of the data. Both indexes prefer $c = 2$ for IRIS at every m except 1.01. This is very consistent with our understanding of these two indexes. When m becomes large, we have pointed out that they will both select $c = 2$—regardless of the data set in question—because of the second limits in (23) and (24). Apparently for Normal-4, large means $m > 1.8$!

ii) *The Xie–Beni Index:* The Xie–Beni index v_{XB}^{FCM} indicates $c = 4$ for Normal-4 over the range 1.01–3.0 for m and over 2–10 for c. And it is perfect, $c = 2$, for all values of c and m on IRIS. This is a very good showing.

iii) *The Extended FCM Xie–Beni Index:* $v_{XB,m}^{\mathrm{FCM}}$ does very nearly as well as v_{XB}^{FCM}, making only one more mistake (at $c = 10$ at $m = 7$ for IRIS). This may be an indicant of its limiting behavior manifested by the choice $c = 10$ for high values of m for both data sets.

iv) *The Fukuyama–Sugeno Index:* $v_{FS,m}^{\mathrm{FCM}}$, which selects the correct value of c only if m lies between 1.5 and 2.8 is the least effective of the five indexes on Normal-4. Its limiting behavior is seen for this data, since it indicates $c = 10$ for both low and high values of m. $v_{FS,m}^{\mathrm{FCM}}$ does a somewhat better job on IRIS, indicating the preferred value of c over the range 2–7.

VI. DISCUSSION AND CONCLUSIONS

We have illuminated the role of model parameters as they affect attempts to validate clusters. Clustering outputs are at the mercy of three things: the data they process, their model parameters, and their algorithmic protocols. We have no control over the data, so when we try to validate outputs of clustering algorithms, it is very important to remember the parameters and the protocols. What our study has shown is that some validity indexes have surprising and sometimes unpredictable dependency on elements of the solution that seem at first glance to be rather unrelated to their job—which is to tell you whether or not to believe the outputs.

Specifically, we have analyzed the role of weighting exponent m in the FCM model as it affects the quality of inferences we can make about the validity of FCM $(U, \mathbf{V})$ pairs produced by any algorithm that attempts to optimize J_m, the fuzzy c-means objective functional. We have seen that, among the indexes tested and for the data sets and protocols used, the Fukuyama–Sugeno measure is much more unreliable, because of its limit properties, than the others. And the same set of experiments suggest that the Xie–Beni index is the most reliable. A useful by-product of our study is this recommendation. Approach FCM $(U, \mathbf{V})$ pairs generated as extrema of the fuzzy c-means model for values of m less than about 1.5 or greater than about 2.5 with even more caution than the level needed for m in [1.5, 2.5]. As with all empirical studies, of course, the next data set tested might suggest otherwise.

How general are the conclusions? We have ignored other indexes—for example, Gunderson's separation coefficient [15], Windham's proportion exponent [12] and uniform data functional [13], and Bensaid's generalization of the Xie–Beni index [20]. The method displayed, however, is quite general. The limit analysis given here can—and should—be applied to any index that is used to evaluate FCM-optimal (U, $\mathbf{V}$) pairs because it is based on necessary conditions for J_m. And more generally, of course, the idea of analyzing the influence of secondary (beyond c) parameters of any clustering algorithm on the validity functions that will be used to evaluate its outputs is very important and should be done whenever possible.

To conclude, we offer this observation. Even if the objects being clustered are well separated into c recognizable subsets, there are many reasons why we may not discover this structure through clustering. For example, the numerical representation of the objects may not possess adequate information to discriminate between clusters of objects. Further, even if the data possess the desired substructure, the algorithm used may not extract it from the data. (For example, an algorithm which looks for hyperspherical clusters will not extract shell type clusters.) Finally, the objects may have structure, the data may represent it, and the algorithm may be capable of finding it, but the appropriate parameters

$$(\underbrace{p_{i1}}_{c}\ p_{i2}, \cdots, p_{ik_i})$$

of the algorithm that yield a successful interpretation of X are never used. And, even if all of these obstacles are met, validity indexes may fail to tell you that the great clusters you have are indeed great! Our goal? Do not give up.

REFERENCES

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

[2] B. S. Everitt, *Graphical Techniques for Multivariate Data.* New York: North Holland, 1978.

[3] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley Interscience, 1973.

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

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

[6] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, 1993.

[7] J. C. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. dissertation, Cornell University, Ithaca, NY, 1973.

[8] ——, "Cluster validity with fuzzy sets," *J. Cybernet.*, vol. 3, no. 3, pp. 58–72, 1974.

[9] ——, "Mathematical models for systematics and taxonomy," in *Proc. 8th Int. Conf. Numerical Taxonomy,* G. Estabrook, Ed., Freeman, San Franscisco, CA, 1975, pp. 143–166.

[10] J. C. Dunn, "Indices of partition fuzziness and the detection of clusters in large data sets," in *Fuzzy Automata and Decision Processes,* M. M. Gupta, Ed. New York: Elsevier, 1976.

[11] J. C. Bezdek, M. Windham, and R. Ehrlich, "Statistical parameters of fuzzy cluster validity functionals," *Int. J. Comp. Infor. Sci.*, vol. 9, no. 4, pp. 232–336, 1980.

[12] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Syst.*, vol. 5, 177–185, 1981.

[13] ——, "Cluster validity for the fuzzy c-means clustering algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 4, no. 4, pp. 357–363, 1982.

[14] M. Roubens, "Fuzzy clustering algorithms and their cluster validity," *European J. Op. Res.*, vol. 10, pp. 294–301, 1982.

[15] R. Gunderson, "Applications of fuzzy ISODATA algorithms to startracker printing systems," in *Proc. 7th Triannual World IFAC Congr.*, 1978, pp. 1319–1323.

[16] X. L. Xie and G. A. Beni, "Validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 3, no. 8, pp. 841–846, 1991.

[17] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Proc. 5th Fuzzy Syst. Symp.*, 1989, pp. 247–250 (in Japanese).

[18] M. Sugeno and T. Yasakawa, "A fuzzy logic based approach to qualitative modeling," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 1, pp. 7–31, 1993.

[19] E. Anderson, "The IRISes of the Gaspe peninsula," *Bull. Amer. IRIS Soc.*, vol. 59, pp. 2–5, 1935.

[20] A. Bensaid, "Improved fuzzy clustering for pattern recognition with applications to image segmentation," Ph.D. dissertation, Univ. of South Florida, Tampa, FL, 1994.