Original paper

Soft Computing 1 (1997) 166—179 © Springer-Verlag 1997

A geometric approach to cluster validity for normal mixtures

J.C. Bezdek, W.Q. Li, Y. Attikiouzel, M. Windham

Abstract We study indices for choosing the correct number of components in a mixture of normal distributions. Previous studies have been confined to indices based wholly on probabilistic models. Viewing mixture decomposition as probabilistic clustering (where the emphasis is on partitioning for geometric substructure) as opposed to parametric estimation enables us to introduce both fuzzy and crisp measures of cluster validity for this problem. We presume the underlying samples to be unlabeled, and use the *expectation-maximization* (EM) algorithm to find clusters in the data. We test 16 probabilistic, 3 fuzzy and 4 crisp indices on 12 data sets that are samples from bivariate normal mixtures having either 3 or 6 components. Over three run averages based on different initializations of EM, 10 of the 23 indices tested for choosing the right number of mixture components were correct in at least 9 of the 12 trials. Among these were the fuzzy index of Xie-Beni, the crisp Davies-Bouldin index, and two crisp indices that are recent generalizations of Dunn's index.

Keywords cluster validity, EM algorithm, generalized Dunn's index, mixture decomposition, normal mixtures, Xie-Beni index

1
Introduction

This paper concerns methods for choosing the best number of components in a set of unlabeled data when the data are drawn from a population whose probability density function is a mixture of p-variate normal distributions [1]. Our objective is to show that non-probabilistic measures of validity can be useful for this problem, so we start by describing its general structure in the context of clustering.

We represent the unlabeled data as $\mathrm{X}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, a set of (n) feature vectors in *feature space* $\Re^p$. The j-th object observed in the process (some physical entity such as a tank, airplane, image pixel, medical patient, stock market report, etc.) has vector $\mathbf{x}_j$ as it's numerical representation; x_{kj} is the k-th characteristic (or *feature*) associated with object j. The most basic idea we need for our discussion is the *class label.* There are four types of class labels: *crisp*, *fuzzy*, *probabilistic and possibilistic.* Let integer c denote the number of classes, $1<c<n$, and define three sets of *label vectors* $\Re^c$ as follows:

$$N_{\mathrm{pc}}=\{\mathbf{p}\in\Re^c\colon \mathrm{p_i}\in[0,1]\ \forall \mathrm{i},\ \mathrm{p_i}>0\ \exists \mathrm{i}\} \tag{1a}$$

$$N_{\mathrm{fc}}=\{\mathbf{p}\in N_{\mathrm{pc}}\colon \sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{p_i}=1\} \tag{1b}$$

$$N_{\mathrm{hc}}=\{\mathbf{p}\in N_{\mathrm{fc}}\colon \mathrm{p_i}\in\{0,1\}\ \forall \mathrm{i}\} \tag{1c}$$

Figure 1 illustrates these three sets for $c=3$. N_{hc} is the canonical (unit vector) basis of $\Re^c$. The i-th vertex of N_{hc}, $\mathbf{e}_\mathrm{i}=(0, 0, \ldots, \underbrace{1}_{\mathrm{i}}, \ldots, 0)^\mathrm{T}$, is the *crisp* label for class i, $1\leqslant i\leqslant c$. The set N_{fc} is a piece of a hyperplane, and is the convex hull of $\mathrm{N_{hc}}$. The vector $\mathbf{p}=(0.1, 0.6, 0.3)^\mathrm{T}$ in Figure 1 is a constrained label vector in N_{f3}; its entries lie between 0 and 1, and sum to 1. There are at least two interpretations for the elements of N_{fc}. If $\mathbf{p}$ comes from a method such as maximum likelihood estimation in mixture decomposition, $\mathbf{p}$ is a (usually posterior) *probabilistic* label, and p_i is interpreted as the probability that, given $\mathbf{x}$, it is in, or came from, class or component i of the mixture [1]. On the other hand, if $\mathbf{p}$ is a label vector for some $\mathbf{x}\in\Re^p$ generated by, say, the fuzzy c-means clustering model [2], $\mathbf{p}$ is a *fuzzy* label for $\mathbf{x}$, and p_i is interpreted as the membership of $\mathbf{x}$ in class i. An important point for this paper is that N_{fc} has the same structure for probabilistic and fuzzy labels.

Finally, the set $N_{\mathrm{pc}}=[0,1]^\mathrm{c}-\{\mathbf{0}\}$ is the unit (hyper)cube in $\Re^c$, *excluding the origin.* Vectors such as $\mathbf{z}=(0.7, 0.2, 0.7)^\mathrm{T}$ in N_{p3} are called *possibilistic* label vectors, and in this case $\mathrm{z_i}$ is interpreted as the possibility that $\mathbf{x}$ is in, or came from, class i. Labels in N_{pc} are produced, e.g., by possibilistic clustering algorithms [3]. Since the constraint $\sum_{\mathrm{i}=1}^{\mathrm{c}} p_\mathrm{i}=1$ for labels in N_{fc} is relaxed to the non-probabilistic condition $\sum_{\mathrm{i}=1}^{\mathrm{c}} p_\mathrm{i}\leqslant c$ for p in N_{pc}, possibilistic methods are not considered in this paper. Note that $N_{\mathrm{hc}}\subset N_{\mathrm{fc}}\subset N_{\mathrm{pc}}$.

Received: *29 July 1997*/*Accepted*: *1 September 1997*

J.C. Bezdek,
Department of Computer Science, University of West Florida
Pensacola, FL 32514 USA

W.Q. Li, Y. Attikiouzel
Department of Electrical and Electronic Engineering,
University of Western Australia Nedlands, Perth
Western Australia, 6009, Australia

M. Windham
Department of Mathematics and Statistics,
University of South Alabama Mobile, AL 36688 USA

Correspondence to: J.C. Bezdek

Research supported by ONR grant #N00014-96-1-0642.

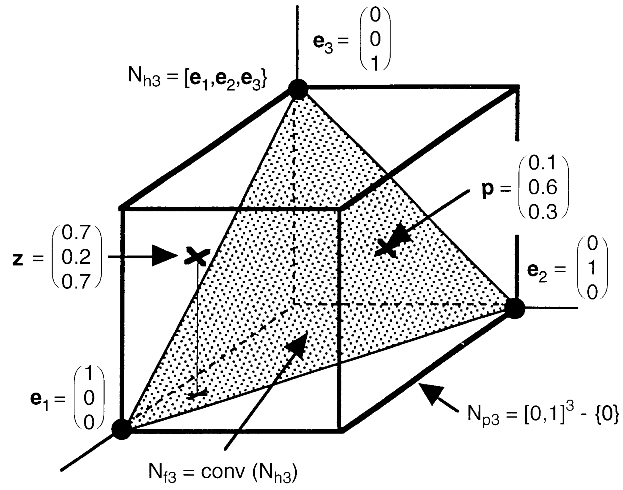
$\mathbf{e}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$
$N_{h3} = [\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3]$
$\mathbf{p} = \begin{pmatrix}0.1\\0.6\\0.3\end{pmatrix}$
$\mathbf{z} = \begin{pmatrix}0.7\\0.2\\0.7\end{pmatrix}$
$\mathbf{e}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$
$\mathbf{e}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$
$N_{p3} = [0,1]^3 - \{0\}$
$N_{f3} = \mathrm{conv}\,(N_{h3})$

Fig. 1. Label vectors in $\mathscr{R}^3$

When X is unlabeled, the assignment of (crisp or fuzzy or probabilistic or possibilistic) label vectors to its elements is called *clustering*[1] (or unsupervised learning). If the labels are crisp, we hope they identify c natural subgroups (clusters) in X. There are several equivalent representations for partitions of data. We find it convenient to define a partition of X as a set of (cn) values $\{u_{ik}\}$ that are arrayed as a ($c \times n$) matrix $U = [\mathbf{U}_1 \ldots \mathbf{U}_k \ldots \mathbf{U}_n] = [u_{ik}]$, where $\mathbf{U}_k$ denotes the k-th *column* of U. The label vectors in (1) can be used to define three types of c-partitions:

$$M_{pcn} = \left\{ U \in \mathfrak{R}^{cn}\text{: } \mathbf{U}_k \in N_{pc} \forall \mathbf{k}\text{: } 0 < \sum_{k=1}^{n} u_{ik} \forall \mathrm{i} \right\} \tag{2a}$$

$$M_{fcn} = \{U \in \mathbf{M}_{pcn}\text{: } \mathbf{U}_k \in \mathbf{N}_{fc} \forall \mathbf{k}\} \tag{2b}$$

$$M_{hcn} = \{U \in \mathbf{M}_{fcn}\text{: } \mathbf{U}_k \in \mathbf{N}_{hc} \forall \mathbf{k}\} \tag{2c}$$

Equations (2a–c) define, respectively, the sets of *possibilistic*, *probabilistic* (or *fuzzy*), and *crisp c-partitions of X*. Each column of U in M_{pcn} (M_{fcn}, M_{hcn}) is a label vector from N_{pc} (N_{fc}, N_{hc}). The constraint $0 < \sum_{k=1}^{n} u_{ik}$ guarantees that the cluster represented by row i of U is not empty. Note that $M_{hcn} \subset M_{fcn} \subset M_{pcn}$. We indicate that U is probabilistic by writing U as $P = [p_{ik}]$, and take p_{ik} as the posterior probability that, given $\mathbf{x}_k$, it came from class i. If U is fuzzy, u_{ik} is the *membership* of $\mathbf{x}_k$ in the i-th fuzzy cluster of X. There are many clustering models, and many algorithms that can be used to look for optimal solutions to a particular model [1, 2, 4].

We represent clustering algorithms as functions $\mathscr{C}: \mathrm{X} \mapsto \mathscr{R}_{\mathscr{C}}$, where $\mathscr{R}_{\mathscr{C}}$ is the range of $\mathscr{C}$. When the output of $\mathscr{C}$ is *just* a partition (the relational clustering model known as single linkage [4], for example), $\mathscr{R}_{\mathscr{C}} = M_{pcn}$. Many clustering algorithms produce outputs besides partitions. The most common example is a second set of parameters called *point prototypes* (or cluster centers) $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$, $\mathbf{v}_1 \in \mathscr{R}^p \forall \mathrm{i}$. For example, the crisp, fuzzy and possibilistic c-means models are defined jointly in the paired variables (U, $\mathbf{V}$), and for these cases, $\mathscr{R}_{\mathscr{C}} = M_{pcn} \times \mathfrak{R}^{cp}$.

The clustering algorithm we will use is the *Expectation-Maximization* (EM) algorithm [1], which seeks local maxima of the likelihood function of X under assumptions to be given in Section 2. This algorithm has for its range the parameter space $\mathscr{R}_{\mathscr{C}_{EM}} = M_{fcn} \times \mathfrak{R}^p \times \mathfrak{R}^{cp} \times PD^{cp}$, where PD^p is the set of all *positive-definite* (*PD*) $p \times p$ matrices. The parameters estimated by $\mathscr{C}_{EM}$ are not all independent. Specifically, estimates $P = [p_{ik}] \in M_{fcn}$ of the (cn) posterior probabilities of the mixture population for samples drawn from normal mixtures are coupled to estimates $(\mathbf{p}, \{\mathbf{m}_i\}, \{S_i\}) \in \mathfrak{R}^p \times \mathfrak{R}^{cp} \times PD^{cp}$ of the priors (π), means $\{\boldsymbol{\mu}_i\}$ and covariance matrices $\{\Sigma_i\}$ of the population parameters through Bayes rule. We will exploit this fact when using fuzzy validity indices to assess the validity of normal mixtures.

Let $\mathscr{P} = \{P_i \in M_{fcn}\text{: } 1 \leqslant \mathrm{i} \leqslant N\}$ denote N different partitions (with or without extra parameters such as $((\mathbf{p}, \{\mathbf{m}_i\}, \{S_i\})$ or $\mathbf{V}$) of a fixed data set X that may arise as a result of

(i) clustering X with one algorithm $\mathscr{C}$ at various values of c; or
(ii) clustering X over other algorithmic parameters of $\mathscr{C}$; or
(iii) applying different $\{\mathscr{C}_j\}$ to X, each with various parameters; or
(iv) all of the above.

Cluster validity is the study (selection or rejection) of which $P_i \in \mathscr{P}$ is "best" in some well defined sense. Visual examination of the algorithmically suggested structure in data is possible only for $p \leqslant 3$, so human validation is specialized to a small (but still important, as for example, in the segmentation of *Magnetic Resonance Images* (MRI) by clustering [5, 6]) subset of the problems that clustering is used for. Moreover, human validation is subjective and to some extent non-repeatable. Consequently, we must often rely on mathematical validation.

Validity functionals $\mathscr{V}: \mathscr{D}_{\mathscr{V}} \mapsto \mathfrak{R}$, $\mathscr{D}_{\mathscr{V}}$ denoting the domain of $\mathscr{V}$, are used to mathematically rank $P_i \in \mathscr{P}$. $\mathscr{D}_{\mathscr{V}}$ is usually (but not necessarily) chosen to match the range of $\mathscr{C}$, $\mathscr{D}_{\mathscr{V}} = \mathscr{R}_{\mathscr{C}}$. When $\mathscr{D}_{\mathscr{V}} = M_{hcn}$, we call $\mathscr{V}$ a *direct measure* because it assesses properties of crisp (real) clusters or subsets in X; otherwise, it is *indirect*.

There are two ways to view $\mathscr{C}$, and hence, two ways to approach the problem of how to define the best partition of X. First, it is possible to regard $\mathscr{C}$ as a *parametric estimation method - P* (and any additional parameters such as $\mathbf{p}$, $\{\mathbf{m}_i\}$, $\{S_i\}$)) are being estimated by $\mathscr{C}$ using X. In this case $\mathscr{V}$ is regarded as a measure of goodness of fit of the estimated parameters (to a true but *unknown* set!). When $\mathscr{D}_{\mathscr{V}} = M_{hcn} \times \underbrace{\text{other parameters,}}_{\text{e.g. } \mathfrak{R}^c \times \mathfrak{R}^{cp} \times \mathrm{PD}^{cp}}$ the test $\mathscr{V}$ performs is still direct, and otherwise, it is still indirect.

The second interpretation of $\mathscr{C}$ is in the sense of exploratory data analysis in unlabeled data. When $\mathscr{V}$ assesses U alone, $\mathscr{V}$ is interpreted as a measure of the quality of U in the sense of partitioning for substructure. Again, if U is crisp, the test is direct, and otherwise, it is indirect.

[1] Many writers call this *classification* . It *is* classification of the vectors in X, accomplished by partitioning X. We prefer to reserve the term classification for the more ambitious undertaking of classifier design, which amounts to partitioning all of $\mathfrak{R}^p$.

There have been countless studies of direct and indirect validity indices for crisp [7–9], probabilistic [10–12] and fuzzy [13–15] clustering methods. Some establish theoretical properties of a particular index; others offer simulations that compare various candidates. The reason they are countless, of course, is that when X is not labeled, the true parameters of any model attempting to represent substructure in it are (and always will be) unknown. Consequently, validity functionals have little chance of being *generally* useful for identifying the "best" solution even within a restricted class of models and/or data processes. More typically, they are relied upon instead to eliminate *badly wrong* solutions, so they are usually used as part of the processing done *prior* to final validation by humans or rule bases. Put another way, since none of us know the right answer for unlabeled data when $p > 3$, there is no set of parameters to measure the fit to, nor structure to visually assess. Nonetheless, there are important applications (such as image segmentation) that will tolerate a "best under these circumstances" solution, so these studies are needed and valuable.

In the investigations we are aware of, the three major types of indices - crisp, fuzzy and probabilistic - never appear together[2]. To our knowledge, there has never been a study that "crossed the boundaries" so to speak, by using, say, a measure of fuzzy cluster validity to assess partitions produced by $\mathscr{C}_{\mathrm{EM}}$. This paper will do that.

One of the methods we propose is that validation of the best number of components in a probabilistic mixture can be studied with direct indices of crisp cluster validity. In order to do this, probabilistic labels in N_{fc} must be transformed into crisp labels. Most often, non-crisp labels ($\mathbf{p}$) are converted to crisp ones ($\mathbf{e}_{\mathrm{i}}$) using the conversion function $\mathbf{H}$: $N_{\mathrm{pc}} \mapsto N_{\mathrm{hc}}$.

$$\mathbf{H}(\mathbf{p}) = \mathbf{e}_{\mathrm{i}} \Leftrightarrow \|\mathbf{p} - \mathbf{e}_{\mathrm{i}}\| \leqslant \|\mathbf{p} - \mathbf{e}_{\mathrm{j}}\| \Leftrightarrow \mathrm{p}_{\mathrm{i}} \geqslant \mathrm{p}_{\mathrm{j}};\ \mathrm{j} \neq \mathrm{i} \tag{3}$$

In (3) $\|*\|$ is the Euclidean distance $\|\mathbf{p} - \mathbf{e}_{\mathrm{i}}\| = \sqrt{(\mathbf{p} - \mathbf{e}_{\mathrm{i}})^{\mathrm{T}}(\mathbf{p} - \mathbf{e}_{\mathrm{i}})}$ on $\mathfrak{R}^{\mathrm{c}}$, and ties are broken arbitrarily. $\mathbf{H}$ simply finds the crisp label vector $\mathbf{e}_{\mathrm{i}}$ in N_{hc} closest to $\mathbf{p}$ and uses it instead of $\mathbf{p}$. Alternatively, $\mathbf{H}$ finds the *maximum coordinate* of $\mathbf{p}$, and assigns the corresponding crisp label to the object $\mathbf{z}$ that $\mathbf{p}$ labels. The rationale for using $\mathbf{H}$ depends on the algorithm that produces $\mathbf{p}$. For example, the justification for computing $\mathbf{H}(\mathbf{p})$ when $\mathbf{p}$ comes from the k-nearest neighbor rule [4] is simple majority voting. If $\mathbf{p}$ is gotten from mixture decomposition, using $\mathbf{H}$ is Bayes rule - label $\mathbf{z}$ by its class of maximum posterior probability. And if the labels are fuzzy, this step is called defuzzification of U by the maximum membership rule. We call this operation the *hardening*, of $\mathbf{p}$ by $\mathbf{H}$. Partitions of X are hardened by applying $\mathbf{H}$ to each column of U in M_{pcn}. In particular, we will apply $\mathbf{H}$ to the columns of posterior probability matrices $P \in M_{\mathrm{fcn}}$ gotten from $\mathscr{C}_{\mathrm{EM}}$, resulting in crisp maximum posterior probability matrices P_{MP}: that are almost always in M_{hcn}. (There is no guarantee that hardening produces a non-degenerate partition (one that has one of more zero rows), but we have never seen this happen.)

2 The Normal Mixture Model and the EM Algorithm for Maximum-Likelihood Estimation

Greek letters are used for population parameters, and Arabic letters for estimates of them. X is assumed to be a set of n observations drawn i.i.d from a mixed population of c p-variate probability distributions with random variables $\{\mathbf{X}_{\mathrm{i}}\}$, that have $\{\pi_{\mathrm{i}}\}$ as their *prior probabilities* (or mixing proportions) and $\{\mathrm{g}(\mathbf{x}|\mathrm{i})\}$as their class-conditional *probability density functions* (PDFs) . The convex combination

$$\mathrm{f}(\mathbf{x}) = \sum_{\mathrm{i}=1}^{\mathrm{c}} \pi_{\mathrm{i}} \mathrm{g}(\mathbf{x}|\mathrm{i}) \tag{4}$$

is itself a PDF that is called a *mixture* of the c components $\{\pi_{\mathrm{i}}\mathrm{g}(\mathbf{x}|\mathrm{i})\}$. Let the *posterior probability* that, given $\mathbf{x}$, $\mathbf{x}$ came from class i, be denoted by $\pi(\mathrm{i}|\mathbf{x})$. Bayes rule relates the elements of (4) to the probabilities $\{\pi(\mathrm{i}|\mathbf{x})\}$ as follows:

$$\pi(\mathrm{i}|\mathbf{x}) = \pi_{\mathrm{i}}\mathbf{g}(\mathbf{x}|\mathrm{i})/\mathrm{f}(\mathbf{x}) \tag{5}$$

At $\mathbf{x}_{\mathrm{k}}$ in $\mathfrak{R}^{p}$ the posterior probability vector $\pi(*|\mathbf{x}_{\mathrm{k}}) = (\pi(1|\mathbf{x}_{\mathrm{k}}), \pi(2|\mathbf{x}_{\mathrm{k}}), \ldots, \pi(\mathrm{c}|\mathbf{x}_{\mathrm{k}})^{\mathrm{T}}$ is in N_{fc}. For $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{n}}\}$, the $\mathrm{c} \times \mathrm{n}$ matrix $\Pi = [\pi(\mathrm{i}|\mathbf{x}_{\mathrm{k}})] = [\pi_{\mathrm{ik}}]$ of posterior probabilities is in M_{fcn}. And Bayes rule relates the elements of Π to the c components of the mixture as $\pi_{\mathrm{ik}} = \pi_{\mathrm{i}}\mathrm{g}(\mathbf{x}_{\mathrm{k}}|\mathrm{i})/\mathrm{f}(\mathbf{x}_{\mathrm{k}})$. When functional forms for the $\{\mathrm{g}(\mathbf{x}|\mathrm{i})\}$in the mixture model are known, but each depends on an unknown vector $\mathbf{q}_{\mathrm{i}}$, we write

$$\mathrm{f}(\mathbf{x}\colon \mathfrak{g}) = \sum_{\mathrm{i}=1}^{\mathrm{c}} \pi_{\mathrm{i}}\mathrm{g}(\mathbf{x}|\mathrm{i};\, \mathbf{q}_{\mathrm{i}}) \tag{6}$$

The mixture density now depends on a *vector of parameters* $\mathfrak{g} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{\mathrm{c}})$. One of the most popular ways to estimate $\mathfrak{g}$ using unlabeled data begins with the *Maximum Likelihood* (ML) model. Given X, which we assume to be identifiable, we form the log-likelihood function $\mathrm{L}(\mathfrak{g}) = \sum_{\mathrm{k}=1}^{\mathrm{n}} \log \mathrm{f}(\mathbf{x}_{\mathrm{k}}\colon \mathfrak{g})$ of the samples and try to maximize it as a function of $\mathfrak{g}$. The case that dominates applications is when every $\mathrm{g}(\mathbf{x}|\mathrm{i}) = \mathscr{n}(\boldsymbol{\mu}_{\mathrm{i}}, \Sigma_{\mathrm{i}})$ is *multivariate normal*, so that component i has the familiar form

$$\mathrm{g}(\mathbf{x}|\mathrm{i}) = \mathrm{e}^{-\frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}_{\mathrm{i}}\|^2_{\Sigma_{\mathrm{i}}-1}} \Big/ (2\pi)^{\frac{\mathrm{p}}{2}} \sqrt{\det \Sigma_{\mathrm{i}}}, \text{ where} \tag{7a}$$

$\boldsymbol{\mu}_{\mathrm{i}} = (\mu_{1\mathrm{i}}, \mu_{2\mathrm{i}}, \ldots, \mu_{\mathrm{pi}})^{\mathrm{T}}$ is the *population mean vector* of class *i*, and

$$\tag{7b}$$

$$\Sigma_{\mathrm{i}}^{-1} = [\mathrm{cov}(\mathrm{X}_{\mathrm{i}})]^{-1} = \begin{bmatrix} \sigma_{\mathrm{i},11} & \sigma_{\mathrm{i},12} & \cdots & \sigma_{\mathrm{i},1\mathrm{p}} \\ \sigma_{\mathrm{i},21} & \sigma_{\mathrm{i},22} & \cdots & \sigma_{\mathrm{i},2\mathrm{p}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\mathrm{i},\mathrm{p}1} & \sigma_{\mathrm{i},\mathrm{p}2} & \cdots & \sigma_{\mathrm{i},\mathrm{pp}} \end{bmatrix}^{-1}. \tag{7c}$$

Σ_{i} is the (positive definite) *population covariance matrix* of class i. The norm in (7a) is the Mahalanobis norm, $\|\mathbf{x} - \mathbf{v}\|_{\Sigma^{-1}} = \sqrt{(\mathbf{x} - \mathbf{v})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \mathbf{v})}$. In this case the unknown parameters for class i to be estimated are $\mathbf{q}_{\mathrm{i}} = (p_{\mathrm{i}}, \mathbf{m}_{\mathrm{i}}, S_{\mathrm{i}})$, estimates of $(\pi_{\mathrm{i}}, \boldsymbol{\mu}_{\mathrm{i}}, \Sigma_{\mathrm{i}})$[3].

[2] This is probably attributable mainly to the idea that we shouldn't mix apples and oranges - but they can make a great fruit salad when combined judiciously!

[3] Alternatively, estimates P for posterior matrix Π can be viewed as the parameters of (6). Equation (9) couples estimates P for Π to the parameters we call $\mathfrak{g}$ here.

First order *necessary* conditions for the *ML* estimators $(p_i, \mathbf{m}_i, S_i)$ of (π_i, μ_i, Σ_i) are well known [16]. Letting $P=[p_{ik}]$ denote the *ML* estimate for the matrix of posteriors, these are

$$p_i = \sum_{k=1}^{n} p_{ik} \Big/ n;\ 1 \leqslant i \leqslant c; \tag{8a}$$

$$\mathbf{m}_i = \sum_{k=1}^{n} p_{ik}\mathbf{x}_k \Big/ \sum_{k=1}^{n} p_{ik};\ 1 \leqslant i \leqslant c; \tag{8b}$$

$$S_i = \sum_{k=1}^{n} p_{ik}(\mathbf{x}_k - \mathbf{m}_i)(\mathbf{x}_k - \mathbf{m}_i)^T \Big/ \sum_{k=1}^{n} p_{ik};\ 1 \leqslant i \leqslant c;\ \text{and} \tag{8c}$$

$$p_{ik} = \frac{p_i g(\mathbf{x}_k|i;(\mathbf{p}_i,\mathbf{m}_i,S))}{f(\mathbf{x}_k|i;(p_i,\mathbf{m}_i,S_i))};\ 1 \leqslant i \leqslant c;\ 1 \leqslant k \leqslant n;$$

$$g(\mathbf{x}_k|i) \equiv \mathscr{n}(\mathbf{m}_i, S_i). \tag{9}$$

Knowing either P or $\mathfrak{g}$ allows us to compute the other set of variables (estimators) through (8) or (9). The basic form of $\mathscr{C}_{EM}$ in the variables $(P, \mathfrak{g})$ uses *alternating optimization* (AO), which takes the form of coupled equations, either

$$P_t = \mathscr{F}_{EM}(\mathfrak{g}_{t-1});\ \mathfrak{g}_t = \mathscr{G}_{EM}(P_t)\ [\mathfrak{g}\text{-initialization}]\text{: or} \tag{10a}$$

$$\mathfrak{g}_t = \mathscr{G}_{EM}(P_{t-1});\ P_t = \mathscr{F}_{EM}(\mathfrak{g}_t)\ [P\text{-initialization}]. \tag{10b}$$

In (10) $\mathscr{F}_{EM}$ is defined by the right side of (9) and $\mathscr{G}_{EM}$ is defined collectively by the right sides of (8). The iterate sequences in (10a) or (10b) are equivalent. Both are exhibited to point out that you can start (initialize) iteration with either P or $\mathfrak{g}$. Specific implementations use one or the other, and properties of either sequence (such as convergence) automatically follow for iteration started at the other set of variables. Much of the general theory of convergence and extrema for alternating optimization (also called *grouped coordinate descent*) is contained in [17–20]. Depending on your point of view, $\mathscr{C}_{EM}$ is either a clustering algorithm (finding a partition P of X for substructure), or a statistical method of parametric estimation (using X to find ML estimates of $(\{(\pi_i, \boldsymbol{\mu}_i, \Sigma_i)\}$ and hence Π). The implementation of $\mathscr{C}_{EM}$ used for the examples in this article is based on (10b) and is given in the Appendix.

AO schemes are usually split gradient descent methods, and so they suffer from several well known numerical problems. The objective function they try to optimize often has multiple local extrema so they can get trapped at undesirable locations (e.g., saddle points), they need good initializations, and they can even exhibit limit cycle behavior for a given data set. $\mathscr{C}_{EM}$ suffers from all these problems. Moreover, when the covariance structure of one or more components of the mixture is arbitrary, $\mathrm{L}(\mathfrak{g})$ does not have a finite maximum [21]. Nonetheless, numerical solutions of system (8–9) are known to produce useful estimators in many real problems so probabilistic clustering (that is, estimates of the matrix P) gotten through this method are popular. Of course, the efficacy of this approach depends entirely on the data satisfying the assumption that it was drawn i.i.d. from mixture (4) with components (7). If this is not the case (and often, it is not), cluster substructure found by $\mathscr{C}_{EM}$ can be very misleading.

The question of validation for estimates produced by $\mathscr{C}_{EM}$ arises two ways. First, as we change c, the estimated parameters obviously change, and so does the quality of the proposed solution from either the clustering or parametric estimation viewpoint. We will call this *validation across c*. And second, with c fixed there may be any number of competitors generated by $\mathscr{C}_{EM}$ as it ranges across different choices for its control parameters (such as initialization and termination criteria). We call this *validation within c*[4]. Most users of clustering algorithms concern themselves primarily with the across c problem, for it is without question more important to be looking in the right solution space than it is to be looking at competitors within the right space. Nonetheless, both problems exist, and validity functionals can be and are used for both. Our primary concern is across c, but our examples also consider the within c problem that can arise with $\mathscr{C}_{EM}$.

3 Validity indices: a melange

We study three kinds of validity indices - probabilistic, fuzzy and crisp. There are as many of each kind as you have time to try, so we rely heavily on previous surveys and comparisons of each kind to choose just a few fuzzy and crisp measures for this study. A short discussion and tabulation of the ones used from each class follows.

3.1 Probabilistic indices

Probabilistic indices are almost always interpreted as measures of goodness of fit, since the mixture model rests on assumptions about the parameters of its components. Bayes rule couples $(p_i, \mathbf{m}_i, S_i) = \mathscr{G}_{EM}(P)$ and $P = \mathscr{F}_{EM}(p_i, \mathbf{m}_i, S_i)$, so the domain of validity functionals is usually taken as $\mathscr{D}_{\mathscr{V}} = M_{fcn}$ or $\mathscr{D}_{\mathscr{V}} = N_{fc} \times \mathfrak{R}^{cp} \times PD^{cp}$ depending on whether P or $(p_i, \mathbf{m}_i, S_i)$ is the focus of attention. Indices in this category fall into one of five broad types: likelihood ratio tests, information-based criteria, Bozdogan's entropic complexity criteria, minimum information ratios, and everything else. All of the indices used here have appeared in previous studies, so we provide only the briefest discussion of each, and then list their formulae in Table 1.

We did not use any likelihood ratio tests because it is well known that the classical regularity conditions for $-2\log\lambda = -2\mathrm{L}(\mathfrak{g}_{c+1}) + 2L(\mathfrak{g}_c)$ do not hold so the ratio fails to have an asymptotic null distribution of χ^2 [22].

Information-based criteria for validation of mixtures apparently first appeared in Wallace and Boulton [23]. The best known criterion of this type is *Akaike's Information Criterion* [24]. These criteria can be recognized by their first term, which is always $(-2\mathrm{L}(\mathfrak{g}))$. Our study calculated Akaike's AIC1 and Bozdogan's modification called AIC3 [25]. A related set of indices are based on Rissanen's *Minimum Description Length* (MDL) principle [26]. We computed MDL, two variants of it

[4] Statistically oriented users may argue that the within c problem is easily resolved by always taking the largest observed maximum of $\mathrm{L}(\mathfrak{g})$. However, it is easy to find examples where this strategy misleads. Indeed, it is a more general curiosity of optimization theory that the global solution to a well-posed problem may be unsatisfactory on any number of other accounts, this being principally due to the mismatch between the process we humans observe, our models of it, and the data we collect to represent it. See p. 220 in [21] for a convincing example of this failure by the c-means objective function.

Table 1. Seventeen probabilistic indices

Index	Equation*	Find
Fitness	$L(\mathfrak{g})=\sum_{k=1}^{n}\log f(\mathbf{x}_k:\mathfrak{g})$	Max
AIC1	$-2L(\mathfrak{g})+2n_{pc}$	Min
AIC3	$-2L(\mathfrak{g})+3n_{pc}$	Min
MDL1	$-2L(\mathfrak{g})+n_{pc}\log n$	Min
MDL2	$-2L(\mathfrak{g})+2n_{pc}\log n$	Min
MDL5	$-2L(\mathfrak{g})+5n_{pc}\log n$	Min
AWE	$-2L(\mathfrak{g})+2n_{pc}(1.5+\log n)$	Min
ICOMP	$-2L(\mathfrak{g})+cp\left(\frac{p+3}{2}\right)\log\left[\frac{\sum_{i=1}^{c}\left\{\frac{tr(S_i)}{p_i}+\frac{tr(S_i^2)}{2}+\frac{(tr(S_i))^2}{2}+\sum_{j=1}^{p}s_{i,jj}^2\right.}{cp\left(\frac{p+3}{2}\right)}\right]$ $-\left\{(p+2)\sum_{i=1}^{c}\log(\det(S_i))-p\sum_{i=1}^{c}\log(np_i)\right\}-cp\log(2n)$	Min
ICOMP1	$\mathrm{ICOMP}+n_{pc}$	Min
ICOMP2	$\mathrm{ICOMP}+n_{pc}(\log(n)+1)$	Min
ICOMPw	$-2L(\mathfrak{g})+\left(\frac{n_{pc}}{2}\right)\log(tr(\hat{F}(\mathfrak{g})^{-1}/n_{pc})-\log(\det(\hat{F}(\mathfrak{g})^{-1})/2$	Min
NEC	$\left(-\sum_{k=1}^{n}\sum_{i=1}^{c}[p_{ik}\log(p_{ik})]\right)\Big/(L(\mathfrak{g})-L(1))$	Min
MIR1	$1-(\lVert\mathfrak{g}_{r+1}-\mathfrak{g}_r\rVert_1/\lVert\mathfrak{g}_r-\mathfrak{g}_{r-1}\rVert_1)$	Max
MIR2	$1-(\lVert\mathfrak{g}_{r+1}-\mathfrak{g}_r\rVert_2/\lVert\mathfrak{g}_r-\mathfrak{g}_{r-1}\rVert_2)$	Max
MIRs	$1-(\lVert\mathfrak{g}_{r+1}-\mathfrak{g}_r\rVert_\infty/\lVert\mathfrak{g}_r-\mathfrak{g}_{r-1}\rVert_\infty)$	Max
ALL1	$(L(\mathfrak{g})-L(1))\ [1-(\lVert\mathfrak{g}_{r+1}-\mathfrak{g}_r\rVert_1/\lVert\mathfrak{g}_r-\mathfrak{g}_{r-1}\rVert_1)]=(L(\mathfrak{g})-L(1))\bullet \mathrm{MIR1}$	Max
ANC1	$(c-1)\ [1-(\lVert\mathfrak{g}_{r+1}-\mathfrak{g}_r\rVert_1/\lVert\mathfrak{g}_r-\mathfrak{g}_{r-1}\rVert_1)]=(c-1)\bullet \mathrm{MIR1}$	Max

$\mathfrak{g}=(\mathbf{q}_1,\mathbf{q}_2,\ldots,\mathbf{q}_c)$; $\mathbf{q}_i=(p_i,\mathbf{m}_i,S_i)\forall i$: n_{pc} is the total number of estimated parameters at value c: $\|\mathfrak{g}\|_1=\sum_{k=1}^{n_{pc}}|\mathfrak{g}_k|$, $\|\mathfrak{g}\|_2\equiv\sqrt{\sum_{k=1}^{n_{pc}}\mathfrak{g}_k^2}$, $\|\mathfrak{g}\|_\infty\equiv\underbrace{\max}_{1\leqslant k\leqslant n_{pc}}\{|\mathfrak{g}_k|\}$: $\hat{F}(\mathfrak{g})$ is estimate (12) of Fisher's information matrix; ALL1 and ANC1 are functions of MIR1

(MDL2 , MDL5) given in [27], and the *Approximate Weight of Evidence* (AWE) [28]. The heuristic is that all of these validity criteria are to be minimized.

The *information complexity* (ICOMP) criterion was first proposed by Bozdogan [29]. It is related to the negative of the Hessian of L ,which is Fisher's information matrix

$$F(\mathfrak{g})=-H_L(\mathfrak{g})=-[\partial^2 L(\mathfrak{g})/\partial\mathfrak{g}_i\partial\mathfrak{g}_j]. \tag{11}$$

We computed four versions of ICOMP: Bozdogan's original ICOMP, two modifications of it called ICOMP1 and ICOMP2 which add a term to ICOMP that depends on the number of parameters being estimated [29], and Cutler and Windham's ICOMPw. ICOMP, ICOMP1 and ICOMP2 are computed directly with estimates from the EM algorithm whereas ICOMPw is ICOMP based on Cutler and Windham's [11] estimate for F at (11), viz.,

$$F(\mathfrak{g})\approx\hat{F}(\mathfrak{g})=\sum_{k=1}^{n}(\nabla L(\mathbf{x}_k;\mathfrak{g}))\ (\nabla L(\mathbf{x}_k;\mathfrak{g}))^T \tag{12}$$

The *Minimum information ratio* (MIR) was introduced by Cutler and Windham [11]. The MIR is the smallest eigenvalue of the matrix $F_C^{-1}F$. Here F_C is the Fisher information matrix for the "classified" model; i.e, F_C based on the joint density for $\mathbf{x}$ and cluster i, $f(\mathbf{x}, i)$. Matrix F is the Fisher information matrix for the marginal density $\mathbf{f}(\mathbf{x})$. So, F_C is actually unknown but one of the eigenvalues of $F_C^{-1}F$ can be estimated [11]. It has been shown that the convergence rate of the EM algorithm is the largest eigenvalue of $I-F_C^{-1}F$, so the MIR can be estimated as 1 minus the EM convergence rate. We studied three forms of the MIR corresponding to using the 1, 2 and sup norms to measure distances between successive estimates of $\mathfrak{g}$. We also computed two variations of the MIR called ALL and ANC by Cutler and Windham [11]. The only probabilistic criterion we tested that is an explicit function of both P and $\mathfrak{g}$ is Celeux's NEC entropy criteria [12].

Table 1 exhibits the formulae and references for the probabilistic indices used in this study. All indices use the sample size n and they all depend on c, the number of components. A few also use p, the dimension of the normal variates. Nine indices explicitly use n_{pc}, the number of parameters being estimated for value c. This is an attempt to guard these indices against monotonicity in c. All indices that use $L(\mathfrak{g})$ are explicit functions of $\mathfrak{g}$ and X.

Notational consistency would require us to write all of the indices in Table 1 as, e.g. $\mathscr{V}_{ICOMP}(\mathfrak{g})$, except NEC, which in our chosen notation would become $\mathscr{V}_{NEC}(P, \mathfrak{g})$. We don't show them this way because the right sides of these indices make the variables they depend upon quite clear. *Notice that all of these indices are indirect*: none of them operate on real clusters in X. This suggests that most researchers with a statistical bias in mixture validation regard outputs of the EM algorithm as solutions of a parametric estimation problem rather than as a clustering of unlabeled data. This may be because many investigators are really interested in using the estimated parameters as the basis for a Bayes classifier rather than simply acquiring labels for the points in X.

3.2 Fuzzy indices

Fuzzy clustering models have unknown parameters such as U, $\mathbf{V}$ or $(U, \mathbf{V})$ as in the c-means models, but underlying distributional assumptions about the process generating the data are usually not made. Consequently, indices of validity for fuzzy partitions are almost always interpreted as measures of partition quality rather than of goodness of fit. Moreover, like the probabilistic indices in Table 1, they are almost always *indirect*.

Many of the early fuzzy indices were simply measures of the "amount of fuzziness" in a fuzzy partition U in M_{fcn} of X. The basic structure of these measures was that $\mathscr{D}_{\mathscr{V}}=M_{fcn}$ even when $\mathscr{R}_{\mathscr{C}}=M_{fcn}\times\{\text{other parameters}\}$. The first indices of this type were the *partition coefficient* $\mathscr{V}_{PC}(U)=-\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^2/n$ and *partition entropy* $\mathscr{V}_{PE}(U)=-\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\log(u_{ik})/n$ of Bezdek [2]. These were rapidly followed by a number of others, including the *uniform data functional*of Windham [14], the fuzziness index of Roubens [30], and so on. Empirical studies show that sometimes indices like these lead to a good partition, and sometimes they don't. This simply confirms that no matter how good (or what type) your index is, there's a data set out there waiting to trick it (and you).

Indirect indices like the partition coefficient suffer from at least three drawbacks. First, they are at best indirectly related to any real clusters in X; second, they ignore additional parameters (such as $\mathbf{V}$) in the range of $\mathscr{C}$; and third, they do not use X itself (cf., $L(\mathfrak{g})$, which does). Xie and Beni [31] defined an index of fuzzy cluster validity that overcomes the second and third problems. This index is a function of X, any fuzzy partition U of X, and any set of cluster centers $\mathbf{V}=\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ in $\mathfrak{R}^p$,

$$\mathscr{V}_{XB}(U, \mathbf{V})=\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^2\|\mathbf{x}_k-\mathbf{v}_i\|^2}{n\left(\underbrace{\min}_{i\neq j}\{\|\mathbf{v}_i-\mathbf{v}_j\|^2\}\right)}. \tag{13}$$

The form of (13) is related historically to the fuzzy c-means model, but that is unimportant for the current work. What is important is that the posterior matrix P from Bayes rule (9) and the fuzzy partition U have the same mathematical structure, and perform similar roles within their models. And further, the centers $\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ from fuzzy c-means or elsewhere play the same role in fuzzy clustering - viz. as locators of central tendency - as the sample means $\mathbf{M}=(\mathbf{m}_1, \ldots, \mathbf{m}_c)$ in (8b) do in probabilistic clustering. These facts allow us to interpret (13) in the context of mixture validation.

Xie-Beni's rationale for using (13) in validating fuzzy clusters was geometric; good clusters should minimize this index by having compact representations (and therefore small numerators) and wide separations (and therefore large denominators). Probabilistic indices draw their rationale from a different source - viz., as measures of goodness of fit. But the data are immune to our choices about models. In view of the mathematical similarities of the basic structure, our idea is simple: why not use indices like (13) for validating mixtures? This is easy to do. With the assignments $P\leftarrow U$ and $\{\mathbf{m}_i\}\leftarrow\{\mathbf{v}_i\}$, Xie-Beni's index becomes

$$\mathscr{V}_{XB}(P, \mathbf{M}; X)=\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}p_{ik}^2\|\mathbf{x}_k-\mathbf{m}_i\|^2}{n\left(\underbrace{min}_{i\neq j}\{\|\mathbf{m}_i-\mathbf{m}_j\|^2\}\right)}. \tag{14}$$

So, this is the general idea. There are literally dozens of fuzzy cluster validity indices in the literature, and Bayes rule together with the common structure for P and U in M_{fcn} allows you to transform every one of them into a measure that might be useful for mixture validation. We think the actual utility of indices obtained this way is very much dependent on the samples at hand. The reason we expect an index like (14) to work is that it measures *geometric properties that can be reasonably expected to exist in samples of normal mixtures* unless the components are very tightly overlapped. In fact, transformation of fuzzy indices as in (14) is not explicitly tied to *normal* mixtures, so there is no reason they might not be equally effective for other types of component densities provided sample geometry can be reasonable expected to match the model underlying the validity functional. What we will not do in this paper is prove that any of these transformed indices measure *statistically* interesting properties of the sample. We suspect that some of them do, but will put this off until we know if the basic idea works at all.

We have chosen for this first study the three fuzzy indices listed in Table 2. The Xie-Beni index at (13) is well regarded in several studies [15]. We will also use the partition coefficient $\mathscr{V}_{PC}$, which was first discussed in the context of this article (as a possible crossover from fuzzy to probabilistic cluster

Table 2. Three fuzzy indices

Index	Equation[1]	Find
PC	$\sum_{k=1}^{n}\sum_{i=1}^{c} p_{ik}^{2}/n$	Max
PE	$-\left(\sum_{k=1}^{n}\sum_{i=1}^{c}[p_{ik}\log(p_{ik})]\right)\Big/\mathrm{n}$	Min
XB	$\sum_{k=1}^{n}\sum_{i=1}^{c} p_{ik}^{2}\lVert\mathbf{x}_k-\mathbf{m}_i\rVert^{2}\Big/\left(\mathrm{n}\left(\underbrace{\min}_{i\neq j}\{\lVert\mathbf{m}_i-\mathbf{m}_j\rVert^{2}\}\right)\right)$	Min

[1] Conversions $P \leftarrow U$ and $\{\mathbf{m}_i\} \leftarrow \{\mathbf{v}_i\}$ have been made from the referenced notation.

validation) by Windham and Cutler [32]. And finally, we will compute the partition entropy $\mathcal{V}_{PE}$ for data set. This index has been used with mixed success since its introduction in 1973. We choose $\mathcal{V}_{PE}$ because it is one term of a two term decomposition for $L(\mathfrak{g})$, and because its functional form appears in several of the probabilistic criteria above (e.g. $\mathcal{V}_{PE}=(L(\mathfrak{g})-L(1))\mathcal{V}_{NEC}$).

3.3 Crisp indices

How many validation methods for crisp partitions are there? Hubert and Arabie [33] began a paper on this topic in 1985 by saying: “We will not try to review this literature comprehensively since that task would require the length of a monograph”. Since it is not feasible to attempt a comprehensive comparison, we have instead chosen just four crisp indices for evaluating the idea that they can be useful in the context of mixture validation: (i) a well known statistically motivated index due to Davies-Bouldin [34] that is a function of the ratio of the sum of within-cluster scatter to between-cluster separation; (ii) a geometric index due to Dunn [35] that measures how compact and well separated the clusters in X are; and (iii) two generalizations of Dunn’s index that have been recently proposed as improvements to it by Bezdek and Pal [36]. These four indices have geometric rationales that should make them useful for data with the structural properties we expect for samples from normal mixtures.

Crisp indices are based on geometric or statistical properties (or both) of the clusters in X, and because the clusters are real subsets of the data, are usually formulated in terms of the subsets rather than U in M_{hcn}. Specifically, let $U \in M_{hcn} \leftrightarrow X=\bigcup_{i=1}^{c}X_i$; $|X_i|=n_i$, $\varnothing = X_i \cap X_j$ for $i \neq j$. When $\mathcal{R}_{\mathcal{C}}=M_{hcn}$ it is always possible to compute the sample mean $\bar{\mathbf{v}}_i=\sum_{\mathbf{x}\in X_i}\mathbf{x}/n_i$ of each X_i, so many crisp indices use both U (or the $\{X_i\}$) and $\bar{\mathbf{V}}=\{\bar{\mathbf{v}}_1,\bar{\mathbf{v}}_2,\ldots,\bar{\mathbf{v}}_c\}$ in their definition.

Davies and Bouldin [34] discussed a crisp index which is a function of the ratio of sums of pairwise within-cluster scatter to between-cluster separation. Since scatter matrices depend on the geometry of the clusters, this index has both a statistical and geometric rationale.

$$\mathcal{V}_{DB,qt}(U,\bar{\mathbf{V}};X)=\left(\frac{1}{c}\right)\sum_{i=1}^{c}\left[\underbrace{\max}_{j,j\neq i}\{(\alpha_{i,t}+\alpha_{j,t})/(\|\bar{\mathbf{v}}_i-\bar{\mathbf{v}}_j\|_q)\}\right]; \tag{15}$$

where $\alpha_{i,t}=(\sum_{\mathbf{x}\in X_i}\|\mathbf{x}-\bar{\mathbf{v}}_i\|^{t}/|X_i|)^{1/t}$. Here t is a positive integer and $\|*\|^{t}$ is the t-th power of the Euclidean norm, while $q \geqslant 1$ defines the Minkowski q-norm of $\mathbf{v}$, $\|\bar{\mathbf{v}}\|_q=(\sum_{j=1}^{p}|\bar{\mathbf{v}}_j|^{q})^{1/q}$. Parameters q and t can be selected independently. Since minimum within-cluster dispersion and maximum between-class separation are both desirable, low values of $\mathcal{V}_{DB,qt}$ are taken as indicants of good cluster structure. Compare this crisp index to (14), the probabilistically interpreted Xie-Beni index. Both indices are built to capture the same type of information about the geometry of substructure in X that samples from normal mixtures usually possess.

Dunn [35] proposed an index based on geometric considerations that has the same basic rationale as $\mathcal{V}_{DB,qt}$ in that both are designed to identify clusters that are compact and well separated. Let S and T be non empty subsets of $\mathfrak{R}^{p}$, and let d: $\mathfrak{R}^{p}\times\mathfrak{R}^{p}\mapsto\mathfrak{R}^{+}$ be any metric. The standard definitions of the *diameter* Δ_1 of S and the *set distance* δ_1 between S and T are

$$\Delta_1(S)=\underbrace{\max}_{\mathbf{x},\mathbf{y}\in S}\{\mathrm{d}(\mathbf{x},\mathbf{y})\}\text{: and} \tag{16}$$

$$\delta_1(S,T)=\underbrace{\min}_{\substack{\mathbf{x}\in S\\ \mathbf{y}\in T}}\{\mathrm{d}(\mathbf{x},\mathrm{y})\}. \tag{17}$$

Dunn defined the *separation index* for the crisp partition $U\leftrightarrow\{X_1,X_2,\ldots,X_c\}$ of X as

$$\mathcal{V}_D(\mathrm{U};\mathrm{X})=\underbrace{\min}_{1\leqslant i\leqslant c}\left\{\underbrace{\min}_{\substack{1\leqslant j\leqslant c\\ j\neq i}}\left\{\frac{\delta_1(\mathrm{X}_i,\mathrm{X}_j)}{\underbrace{\max}_{1\leqslant k\leqslant c}\{\Delta_1(\mathrm{X}_k)\}}\right\}\right\}. \tag{18}$$

The quantity $\delta_1(X_i,X_j\}$ in the numerator of $\mathcal{V}_D$ is analogous to $\|\bar{\mathbf{v}}_i-\bar{\mathbf{v}}_j\|_q$ in the denominator of $\mathcal{V}_{DB,qt}$; $\delta_1(X_i,X_j)$ is a measure of the distance between clusters that operates directly on the points in the clusters, whereas $\|\bar{\mathbf{v}}_i-\bar{\mathbf{v}}_j\|_q$ uses the distance between their cluster centers for the same purpose. The use of $\Delta_1(\mathrm{X}_k)$ in the denominator of $\mathcal{V}_D$ is analogous to $\alpha_{k,q}$ in the numerator of $\mathcal{V}_{DB,qt}$ both are measures of the scatter volume for cluster X_k. Thus, extrema of $\mathcal{V}_D$ and $\mathcal{V}_{DB,qt}$ share roughly the same geometric objective: maximizing intercluster distances whilst minimizing intracluster distances. Since the measures of separation and compactness in $\mathcal{V}_D$ occur “upside down” from their appearance in $\mathcal{V}_{DB,qt}$, *large* values of $\mathcal{V}_D$ correspond to good clusters. Hence, the number of clusters that *maximizes* $\mathscr{V}_D$ is taken as the most valid partition.

Bezdek and Pal [36] observed that a major problem with $\mathscr{V}_D$ is that both its numerator and denominator can be greatly influenced by a few noisy points (that is, outliers to the main cluster structure) in *X*. This in turn causes $\mathscr{V}_D$ to be rather sensitive to what can be a very small minority in the data. To ameliorate this they proposed several ways to generalize $\mathscr{V}_D$ by using other definitions for the diameter of a set at (16) or the distance between sets at (17).

Let Δ_b be any positive semi-definite (*diameter*) function on $P(\mathfrak{R}^p)$, the power set of $\mathfrak{R}^p$. And let δ_a denote any positive semi-definite, symmetric (*set distance*) function on $P(\mathfrak{R}^p)\times P(\mathfrak{R}^p)$. The general form of (18) using δ_b and Δ_a is:

$$\mathscr{V}_{\delta_a\Delta_b}(U;X)=\underbrace{\min}_{1\leqslant i\leqslant c}\left\{\underbrace{\min}_{\substack{1\leqslant j\leqslant c\\ j\neq i}}\left\{\frac{\delta_a(X_i,X_j)}{\underbrace{\max}_{1\leqslant k\leqslant c}\{\Delta_b(X_k)\}}\right\}\right\}\doteq\mathscr{V}_{ab}. \tag{19}$$

Using 6 δ's and 3 Δ's in (19), including $\mathscr{V}_D=\mathscr{V}_{11}$, Bezdek and Pal concluded that the most reliable combinations were the indices they called the *generalized Dunn's indices* (GDIs) $\mathscr{V}_{33}$ and $\mathscr{V}_{63}$ that are shown in Table 3.

$\mathscr{V}_{33}$ and $\mathscr{V}_{63}$ preserve the geometric rationale of (18), but reduce the brittleness of $\mathscr{V}_D$ to noisy points by using set distances and a diameter that are functions of all the points in *S* and *T*. For example, $\delta_6(S,T)$ is the Hausdorff distance between *S* and *T*, and is a measure of the intercluster distance between *S* and *T*. The diameter $\Delta_3(S)=2\left(\sum_{x\in S} d(\mathbf{x},\bar{\mathbf{v}}_S)/|S|\right)$ used by both of these indices is the average distance from the cluster mean $\bar{\mathbf{v}}_S$ of *S* to the points in cluster *S*. This diameter function is relatively insensitive to the addition or deletion of a few aberrant points (provided the number of points is not too small), and is a measure of the scattering volume of cluster *S*.

In summary, all four of the crisp indices in Table 3 are essentially functions of ratios of pairwise cluster separation to individual scatter volumes.

How will we apply these indices to the mixture decomposition validation problem? First run $\mathscr{C}_{EM}$ on *X* to obtain *P* at (9); apply **H** at (3) to the columns of P to get the crisp maximum probability partition P_{MP}, compute $\bar{\mathbf{v}}_i=\sum_{x\in X_i}\mathbf{x}/n_i$, $1\leqslant i\leqslant c$ (when used by the index), and finally, compute the index itself. This is even further afield from probabilistic cluster validity than the idea in section 3.B, but again, our rationale is that the geometry of the data is independent of any assumptions we care to make about models that match it well. All of the crisp indices tested here measure properties that we expect normal mixtures to possess. We think you will be surprised at the computational results.

Table 3. Four crisp indices

Index	Equation[1]	Find
DB	$\mathscr{V}_{DB,qt}(U,\bar{\mathbf{V}};X)=\left(\frac{1}{c}\right)\sum_{i=1}^{c}\left[\underbrace{\max}_{j,j\neq i}\left\{(\alpha_{i,t}+\alpha_{j,t})/(\lVert\bar{\mathbf{v}}_i-\bar{\mathbf{v}}_j\rVert_q)\right\}\right]$; where	Min
	$\alpha_{i,q}=\left(\frac{1}{\lvert X_i\rvert}\sum_{x\in X_i}\lVert\mathbf{x}-\bar{\mathbf{v}}_i\rVert_q\right)^{1/t}$; $\bar{\mathbf{v}}_i=\sum_{x\in X_i}\mathbf{x}/n_i$, $q=t=2$ in examples.	
Dunn	$\mathscr{V}_D(U;X)=\underbrace{\min}_{1\leqslant i\leqslant c}\left\{\underbrace{\min}_{\substack{1\leqslant j\leqslant c\\ j\neq i}}\left\{\frac{\delta_1(X_i,X_j)}{\underbrace{\max}_{1\leqslant k\leqslant c}\{\Delta_1(X_k)\}}\right\}\right\}.$	Max
GDI33	$\underbrace{\min}_{1\leqslant i\leqslant c}\left\{\underbrace{\min}_{\substack{1\leqslant j\leqslant c\\ j\neq i}}\left\{\frac{\delta_3(X_i,X_j)}{\underbrace{\max}_{1\leqslant k\leqslant c}\{\Delta_3(X_k)\}}\right\}\right\};\quad\left\{\begin{array}{l}\delta_3(S,T)=\delta_{avg}(S,T)=\frac{1}{\lvert S\rvert\lvert T\rvert}\sum_{\substack{x\in S\\ y\in T}}d(\mathbf{x},\mathbf{y})\\ \Delta_3(S)=2\left(\frac{\sum_{x\in S}d(\mathbf{x},\bar{\mathbf{v}}_s)}{\lvert S\rvert}\right),\ \bar{\mathbf{v}}_s=\frac{1}{\lvert S\rvert}\sum_{x\in S}\mathbf{x}\end{array}\right\}$	Max
GD163	$\underbrace{\min}_{1\leqslant i\leqslant c}\left\{\underbrace{\min}_{\substack{1\leqslant j\leqslant c\\ j\neq i}}\left\{\frac{\delta_6(X_i,X_j)}{\underbrace{\max}_{1\leqslant k\leqslant c}\{\Delta_3(X_k)\}}\right\}\right\};\quad\left\{\begin{array}{l}\delta_6(S,T)=\max\{\delta(S,T),\delta(T,S)\},\\ \delta(S,T)=\underbrace{\max}_{\mathbf{x}\in S}\left\{\underbrace{\min}_{\mathbf{y}\in T}\{d(\mathbf{x},\mathbf{y})\}\right\}\\ \delta(T,S)=\underbrace{\max}_{\mathbf{y}\in T}\left\{\underbrace{\min}_{\mathbf{x}\in S}\{d(\mathbf{x},\mathbf{y})\}\right\}.\end{array}\right\}$	Max

[1] S, T are crisp sets: $P_{MP}\in M_{hcn}\leftrightarrow X=\bigcup_{i=1}^{c}X_i$; $\lvert X_i\rvert=n_i$; $\varnothing=X_i\cap X_j$ for $i\neq j$. d: $\mathfrak{R}^p\times\mathfrak{R}^p\rightarrow\mathfrak{R}^c$ is any metric $\mathfrak{R}^p$.

Table 4. Data sets and sample sizes

c	Priors $\{\pi_i\}$	Means $\{\mu_i\}$	S: $\Sigma_i = 0.5\mathrm{I}$	M: $\Sigma_i = \mathrm{I}$	O: $\Sigma_i = 2\mathrm{I}$
3	$\pi_i = 1/3$	r = 2	SA3: n = 1,000 SB3: n = 10,000	MA3: n = 1,000 MB3: n = 10,000	OA3: n = 1,000 OB3: n = 10,000
6	$\pi_i = 1/6$	r = 3	SA6: n = 2,000 SB6: n = 20,000	MA6: n = 2,000 MB6: n = 20,000	OA6: n = 2,000 OB6: n = 20,000

4 The data and computing protocols

It is clear that validity indices depend in rather unpredictable ways on many things besides X, including parameters c, p, n and n_{pc}, the number of parameters being estimated. There are many ways to generate samples, and the sampling scheme chosen certainly influences the quality of inferences that can be based on simulations using the samples drawn.

The data The selection of data sets for our study is motivated by attempts to segment e.g., MR images with the unsupervised EM algorithm [5]. Many clustering models produce good segmentations of medical images [37, 39], but as always, the issue of how many tissue classes to accept (i.e., what is the best value for c?) is a major part of the overall problem. In this application, the minimum number of (tissue) classes is about $c=5$, and the usual minimum number of samples is $n = 256^2 = 65{,}536$ pixel vectors in $\Re^3$. Windham and Cutler state that for c much larger than 3, all validity indices tend to fail [11]. Since the practical application we have in mind requires validation for c ranging from about 5 to 10, the samples for our study were drawn from mixtures of $p=2$ dimensional[5] Gaussian PDFs having either $c=3$ or $c=6$ components. A total of six data sets were generated for each of these two cases, resulting in 12 data sets. Table 4 summarizes the model parameters used for data generation.

The priors were fixed at $1/c$, and the means were distributed on rays separated by equal angles, centered equal distances from the origin at radii r = 2 or 3 according as $c=3$ or 6. Thus, for example, at $c=3$ there are three means on lines separated by 120°, each 2 units from $(0, 0)^T$. We used three sets of covariance matrices, shown in Table 4 as cases **S**, **M** and **O**, which stand for **S***eparated*, **M***edium* and **O***verlapping*. The covariance matrices for these cases were scalar multiples of the identity (I) with diagonals (variances) of 0.5, 1 and 2 respectively, and were equal for the components of each mixture.

To see what the test data looked like, we scatterplotted all 12 data sets. For the **O** data sets with variances of 2 for each component, visual examination does not reveal well defined mixture substructure; a scatterplot of the 20,000 points from OB6, is just a big spot. The **S** data sets did show visual clusters at both $c=3$ and $c=6$, as expected.

Initialization of $\mathscr{C}_{\mathrm{EM}}$ Equations (10) show that initialization can be done by choosing either P_0 in M_{fcn}, or equivalently, $(\mathbf{p}_0, \{\mathbf{m}_{i0}\}, \{S_{i0}\})$. We chose to initialize on P_0. This matrix can be chosen randomly but many authors prefer to start $\mathscr{C}_{\mathrm{EM}}$ near a possible solution by processing X with another clustering algorithm first. For example, the sequential competitive learning model called the dog-rabbit algorithm [40] has been used to partition X by U_f in M_{hcn} first. Subsequently, $P_0 \leftarrow U_f$ provides an initialization for $\mathscr{C}_{\mathrm{EM}}$.

A batch clustering algorithm, *fuzzy c-means* (FCM), can also provide good initializations for the EM algorithm *for normal mixtures* [41–43]. FCM terminates at U_f in M_{fcn}, and again, $P_0 \leftarrow U_f$ initializes $\mathscr{C}_{\mathrm{EM}}$. We used FCM as specified in the Appendix for this purpose. However, FCM is itself an AO algorithm for solving a constrained least squares optimization problem, so it also needs initialization. For this we chose the diagonal of the hyperbox $hb(\mathbf{m}, \mathbf{M})$ with corners $\mathbf{m}$ and $\mathbf{M}$ whose components are

$$m_j = \underbrace{\min}_{k} \{x_{jk}\} \colon j = 1, 2, \ldots, p, \tag{20a}$$

$$M_j = \underbrace{\max}_{k} \{x_{jk}\} \colon j = 1, 2, \ldots, p. \tag{20b}$$

as the basis for three initializations of FCM. We used the standard initialization, which is to choose c cluster centers $\{\mathbf{v}_{i0}\}$ that are equally distributed along the line segment that connects $\mathbf{m}$ and $\mathbf{M}$,

$$\mathbf{v}_{i,0} = \mathbf{m} + \left(\frac{\mathrm{i}-1}{\mathrm{c}-1}\right)(\mathbf{M}-\mathbf{m}),\ \mathrm{i} = 1, 2, \ldots, \mathrm{c} \tag{21}$$

We also used two different sets of $\{\mathbf{v}_{i0}\}$'s that were selected randomly from the diagonal of $\mathrm{hb}(\mathbf{m}, \mathbf{M})$. In each case FCM was terminated at $(U_f, \mathbf{V}_f)$ when the maximum absolute difference in the components of $\{\mathbf{v}_i\}$ between two consecutive iterations was less that 0.01. Each set of EM runs discussed below was repeated three times for these three initializations, taking $P_0 \leftarrow U_{\mathrm{final,FCM}}$ to start $\mathscr{C}_{\mathrm{EM}}$.

Termination of $\mathscr{C}_{\mathrm{EM}}$ Various methods for terminating $\mathscr{C}_{\mathrm{EM}}$ are discussed in the literature. Many terminate it by comparing successive estimates of either P or $(\mathbf{p}, \{\mathbf{m}\}, \{S\})$ to some small number ε. A more traditional method of termination is by comparing the absolute difference of successive values of the total log-likelihood $L(\mathfrak{g})$ to cutoff criterion ε. We tested both

[5] Because multispectral image data often are of dimension $p \geqslant 3$, samples from $p \geqslant 3$-dimensional mixtures would be more desirable. Our choice here strikes a balance between what is desirable and what is computationally realistic.

criteria in our examples, and found very little difference in the final results. For example, when the increment in $L(\mathfrak{g})$ is less than 10^{-4}, the change to estimates of either set of parameters is about the same. In our simulations $\mathscr{C}_{\mathrm{EM}}$ was terminated when $E_t = |L(\mathfrak{g}_t) - L(\mathfrak{g}_{t-1})|$ was less than $\varepsilon = 10^{-3}$ for at least 5 iterations or iterate limit T = 300, whichever came first. With these criteria the *maximum* change suffered by any estimated parameter was about 10^{-4}.

5 Simulation results

For each data set listed in Table 4 we ran $\mathscr{C}_{\mathrm{EM}}$ three times – once for each initialization by an output from FCM – for each c ranging from 2 to 6 for the $c = 3$ component samples, and from 4 to 8 for the c = 6 component samples. To see how the summary tables are derived, we show the value of the DB index for the *best run* (= largest value of $L(\mathfrak{g})$ in each set of three) in

Table 5. Values of the Davies-Bouldin index

c	SA3	MA3	OA3	SB3	MB3	OB3	SA6	MA6	OA6	SB6	MB6	OB6
DB Values for the largest value of L($\mathfrak{g}$)												
2	**0.91**	**1.01**	**1.07**	**0.91**	**1.00**	**1.11**						
3	*1.04*	*1.26*	*1.35*	*1.04*	*1.26*	*1.47*						
4	1.11	1.13	1.43	1.17	1.67	1.35	0.88	1.03	1.17	0.92	1.05	1.11
5	1.40	1.26	1.36	1.40	2.91	1.42	0.79	0.95	1.10	0.79	0.95	**1.09**
6	1.44	1.36	1.15	1.66	3.18	1.35	***0.71***	***0.92***	***1.08***	***0.71***	***0.92***	***1.10***
7							0.86	1.02	1.24	0.88	1.09	1.18
8							1.00	1.09	1.21	1.06	1.22	1.12
Average DB values for three initializations												
2	1.29	1.44	1.52	1.28	1.40	1.58						
3	***0.90***	***1.12***	*1.38*	***0.90***	***1.13***	*1.35*						
4	1.13	1.17	1.42	1.13	1.61	1.38	0.89	1.02	1.15	0.89	1.03	1.13
5	1.38	1.25	1.44	1.39	2.01	1.39	0.78	0.94	1.09	0.78	0.96	**1.09**
6	1.43	1.40	**1.17**	1.57	2.53	**1.30**	***0.70***	***0.92***	***1.08***	***0.70***	***0.91***	*1.10*
7							0.87	1.05	1.25	0.89	1.11	1.19
8							1.00	1.11	1.30	1.06	1.23	1.16

Table 6. Summary of the best (= largest maxima of L($\mathfrak{g}$)) results on 12 data sets

Data Set→	SA3	MA3	OA3	SB3	MB3	OB3	SA6	MA6	OA6	SB6	MB6	OB6	# Corr.
Probabilistic indices													
AIC1	✓	✓	+	✓	✓	✓	+	✓	–	✓	✓	✓	9
AIC3	✓	✓	+	✓	✓	✓	✓	✓	–	✓	✓	✓	10
MDL	✓	✓	–	✓	✓	✓	✓	✓	–	✓	✓	–	9
MDL2	✓	✓	–	✓	✓	✓	✓	–	–	✓	✓	–	8
MDL5	✓	–	–	✓	✓	–	✓	–	–	✓	✓	–	6
AWE	✓	✓	–	✓	✓	✓	✓	–	–	✓	✓	–	8
ICOMP	+	✓	+	+	+	+	+	+	+	+	✓	+	2
ICOMP1	✓	✓	+	✓	+	✓	✓	✓	✓	✓	✓	+	9
ICOMP2	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	–	10
ICOMPw	+	✓	+	✓	✓	✓	✓	✓	✓	✓	✓	+	9
NEC	✓	✓	–	✓	✓	–	✓	–	–	–	–	–	5
MIR1	✓	–	–	–	–	–	–	–	–	✓	✓	–	3
MIR2	✓	–	–	–	–	–	–	–	–	✓	✓	–	3
MIRs	✓	–	–	–	–	+	–	–	–	✓	✓	–	3
ALL	✓	✓	–	✓	✓	–	–	–	–	✓	✓	–	6
ANC	✓	✓	–	✓	✓	+	–	–	–	✓	✓	–	6
Fuzzy indices													
PC	✓	–	–	✓	–	–	–	–	–	–	–	–	2
PE	✓	–	–	✓	–	–	–	–	–	–	–	–	2
XB	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	–	10
Crisp indices													
DB	–	–	–	–	–	–	✓	✓	✓	✓	✓	–	5
DI	✓	✓	–	–	✓	–	+	–	–	–	–	+	3
GDI33	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	–	10
GDI63	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	–	10

the upper half of Table 5. The lower half of Table 5 contains the *average value* of the DB index over the outputs of the three runs on each data set. The minimum value for each column in the upper and lower halves of Table 5 is the value of c recommended by DB, while the italicized cells in Table 5 correspond to the actual (desired) value of c for each sample.

Comparing bold-face values to italicized cells shows that for the best values, DB underestimated $c=3$ in all 6 attempts. For the $c=6$ samples, DB correctly identified c for the first 5 sets, and underestimated the number of components in **O**B6. Thus, the crisp DB index yielded 7 underestimates and 5 correct values in 12 tries on partitions that had the largest value of $L(\mathfrak{g})$.

Turning to the average values in the lower half of Table 5, we find 9 correct identifications of the number of components, two overestimations, and one underestimation. By using averages of three runs then, DB improves from 5 to 9 successes. The improvement occurs on 4 of the 6 data sets that have $c=3$, all of which move from $c=2$ at the largest $L(\mathfrak{g})$ to $c=3$ over three runs. This indicates that the EM algorithm is definitely terminating at somewhat different estimates when begun from different initializations. Since many other indices showed less dramatic improvements on moving from best to averages, it is probably also the case that the DB index is particularly sensitive to changes in crisp labels that occur during the application of (3) to matrix P.

Tables 6 and 7 contain summaries of the values of c indicated by all 23 indices using analyses similar to that just offered for Table 5. In these tables, "✓" means that the correct value for c was chosen by the index; "+" means that the index overestimated *c*, and "−" means that the index underestimated *c*.

First we discuss the best single run (of three) for each index, where again, best means that we find the run at each *c* corresponding to the largest maximum of $L(\mathfrak{g})$ over three initializations. Then, all 23 indices are computed and used to select c from the choices over its range for each of the 12 data sets. Table 6 shows the best run statistics for the 23 indices. Two probabilistic indices, AIC3 and ICOMP2 scored 10 successes, as did the crisp indices GDI33 and GDI63, and as did the fuzzy XB index. Four indices (AIC1, MDL, ICOMP1, ICOMPw) identified 9 of 12. As a crude first observation then, it does appear that all three types of indices (probabilistic, fuzzy and crisp) can be useful for validation in the context of normal mixtures: 3 of the top 8 in this first trial were not probabilistic. At the other end of the spectrum, the three MIR indices only had 3 successes in 12 tries, while ICOMP and the fuzzy indices PC and PE and had only 2 successes in 12 tries.

As expected, the indices generally have the best success with the well separated (**S** series) data, and the least success with the overlapping (**O** series) data. The best overall success is for **S**A3, (column 2 of Table 6), where all but 3 of the

Table 7. Summary of the (3 run) averages on 12 data sets

Data Set→	**S**A3	**M**A3	**O**A3	**S**B3	**M**B3	**O**B3	**S**A6	**M**A6	**O**A6	**S**B6	**M**B6	**O**B6	# Corr.
Probabilistic indices													
AIC1	✓	✓	+	✓	✓	✓	+	✓	−	✓	✓	✓	9
AIC3	✓	✓	✓	✓	✓	✓	✓	✓	−	✓	✓	✓	11
MDL	✓	✓	−	✓	✓	✓	✓	✓	−	✓	✓	−	9
MDL2	✓	✓	−	✓	✓	✓	✓	−	−	✓	✓	−	8
MDL5	✓	−	−	✓	✓	−	✓	−	−	✓	✓	−	6
AWE	✓	✓	−	✓	✓	✓	✓	−	−	✓	✓	−	8
ICOMP	+	✓	+	+	+	+	+	✓	+	✓	✓	+	4
ICOMP1	✓	✓	+	✓	✓	✓	✓	✓	✓	✓	✓	+	10
ICOMP2	✓	✓	✓	✓	✓	✓	✓	✓	−	✓	✓	−	10
ICOMPw	✓	✓	+	✓	✓	✓	+	✓	✓	✓	✓	✓	10
NEC	✓	✓	−	✓	✓	−	✓	−	−	✓	−	−	6
MIR1	✓	−	−	−	−	−	−	−	−	−	✓	✓	3
MIR2	✓	−	−	−	−	−	−	−	−	✓	✓	−	3
MIRs	✓	−	−	−	−	−	−	−	−	✓	✓	−	3
ALL	✓	✓	−	✓	✓	−	✓	−	−	✓	✓	✓	8
ANC	✓	✓	−	✓	✓	−	✓	−	−	✓	✓	✓	8
Fuzzy indices													
PC	✓	−	−	✓	−	−	−	−	−	−	−	−	2
PE	✓	−	−	✓	−	−	−	−	−	−	−	−	2
XB	✓	✓	✓	✓	✓	✓	✓	✓	−	✓	✓	−	10
Crisp indices													
DB	✓	✓	+	✓	✓	+	✓	✓	✓	✓	✓	−	9
DI	✓	✓	−	−	✓	−	✓	−	−	−	−	+	4
GDI33	✓	✓	✓	✓	✓	✓	✓	✓	−	✓	✓	−	10
GDI63	✓	✓	✓	✓	✓	✓	✓	✓	−	✓	✓	−	10

23 indices chose $c=3$. This is not surprising since SA3 is the smallest, most well separated, lowest number of clusters data set. At the other extreme, the next to last column of Table 6 shows that only 2 of the 23 indices correctly assessed the best value EM partition of OB6, which has $c=6$ clusters of 20,000 tightly overlapped points. On one hand, this tends to confirm Windham and Cutler's opinion about expectations for validity functionals as c increases. On the other hand, the number of successes for SA6 ($n=2{,}000$) increases from 13 to 18 when SB6 ($n=20{,}000$) is processed; and increases to 19 successes for the more tightly mixed data set MB6. This suggests that higher values of c can be detected by many of these indices provided that the mixture is reasonably well separated, and that the number of samples is adequate.

From Table 6 we can also conclude that, with the exception of AIC1 and ICOMP, when indices failed they tended to *underestimate* the number of components about five times more often than to overestimate this number. In the probabilistic index group, AIC1 and MDL tendencies are consistent with previous studies – AIC1 tends to overestimate and MDL tends to underestimate.

Table 7 contains the results of using the *average index* of three runs based on different initializations of $\mathscr{C}_{\mathrm{EM}}$ over each data set. Table 8 compares the successes from Tables 6 and 7, sorted in descending order on the successes from Table 7, and its last column shows the change in each index from best to average. For about 40% of the indices tested (9 of 23) , the number of successful identifications of the correct value of c was higher than the number of successes for the largest single run maximum of $L(\mathfrak{g})$. The most dramatic increase is shown by the crisp Davies-Bouldin index, which jumps four places from 5 to 9. From this we tentatively conclude that using averages improves the success rate of many validity functionals. The runs conducted here hardly warrant stronger statements; averages of many runs would be needed to say anything more assertive than this.

Table 8. Number of successes: best versus averages on 12 data sets

Index	Type	Best	Ave.	Change
AIC3	Prob.	10	11	1
ICOMP1	Prob.	9	10	1
ICOMP2	Prob.	10	10	0
ICOMPw	Prob.	9	10	1
XB	Fuzzy	10	10	0
GDI33	Crisp	10	10	0
GDI63	Crisp	10	10	0
AIC1	Prob.	9	9	0
MDL	Prob.	9	9	0
DB	Crisp	5	9	4
MDL2	Prob.	8	8	0
AWE	Prob.	8	8	0
ANC	Prob.	6	8	2
ALL	Prob.	6	8	2
NEC	Prob.	5	6	1
MDL5	Prob.	6	6	0
MIR2	Prob.	3	3	0
ICOMP	Prob.	2	4	2
DI	Crisp	3	4	1
MIRs	Prob.	3	3	0
PE	Fuzzy	2	2	0
PC	Fuzzy	2	2	0
MIR1	Prob.	3	3	0

Using averages pushed the AIC3 criterion of Bozdogan to the top of the list. Three of the four ICOMP indices, the XB index and both GDI indices were just behind it with 10 successes, and 3 more indices had 9 right in 12 tries. Of the ten indices in Table 7 that had 9 or more successes, four were non-probabilistic: the fuzzy Xie-Beni index, the Davies-Bouldin index, and both generalized Dunn's indices. This supports our conjecture that non-probabilistic indices can be as effective as probabilistically based measures for assessing the number of components in a normal mixture when the samples possess geometric structure that is favored by these indices. Finally, we note that the most successful indices seem to be the ones that use all the information at their disposal – that is, the data X, the matrix P and the centers $\mathbf{M}$. This is most evident in the poor showing of the two fuzzy indices PE and PC, which use only the matrix P to decide about the best number of clusters. Dunn's index uses X and P but not $\mathbf{M}$; it does slightly better than PE and PC, but is also near the bottom of the list.

Notice that all three MIR indices had 3 successes in both the best and average cases. MIR is known to be somewhat unreliable with poorly separated data and experiments have shown that it is easily fooled by symmetry in the data. The latter is evidenced by the fact that two and three clusters were common answers in the 6 cluster examples.

6
Discussion

Our main objective in this article was to establish empirically that an index of cluster validity can be reasonably expected to work when the structure in the data are favorable to its underlying rationale, regardless of the philosophical bias of the index. Tables 6 and 7 support this assertion. Crisp and fuzzy validity indices that assess *geometric properties* of partitions that match the expected structure in samples from mixtures of normal distributions can be as effective as measures that have some statistical basis for estimating the number of components. In particular, measures that assess central tendency and dispersion of subsets of the data can be expected to succeed.

Since other studies often show behavior for a particular index that is quite the opposite of the performance seen here (cf. [11]), the data we have used are an obvious and natural limitation to any far-reaching generality about our conclusions. Our examples have also shown that many indices can be reliable for relatively large values of c if the mixture is not too highly overlapped and enough samples are used.

What we have NOT shown is a theoretical connection between any of the non-probabilistic indices and the distributional assumptions about normal mixtures that were used to generate the data. This would be a logical and desirable next step for the Xie-Beni and generalized Dunn's indices. The main tool in any such analysis will probably be Baye's rule, and the results may depend heavily on the assumption of normal components. Another interesting study would be a "converse" investigation – which probabilistic indices could be used to validate crisp and fuzzy clusters? We have little doubt that several of the measures listed in Table 1 will be useful for this purpose, and hope to base a future investigation on this conjecture.

Appendix 1
The EM Algorithm for Normal Mixtures [1]

```
Store     Unlabeled Object Data $X \in \Re^p$

Pick      ● number of clusters:         $1 < c < n$
          ● max. # of iterations:       $T = 300$
          ● termination measure:        $E_t = |L(\mathfrak{g}_t) - L(\mathfrak{g}_{t-1})|$
          ● termination criterion:      $\varepsilon = 0.001$

Get       ● initial partition: $P_0 \in M_{fcn}$ (from FCM below)

Iterate   $0 \leftarrow t$
          DO UNTIL ($t = T$ or $E_t \leqslant \varepsilon$ (for 5 iterations))
             $\mathfrak{g}_t = \mathscr{G}_{EM}(P_t)$      where $\mathscr{G}_{EM}(P_t)$      is (8a–c)
             $P_t = \mathscr{F}_{EM}(\mathfrak{g}_{t-1})$   where $P_t = \mathscr{F}_{EM}(\mathfrak{g}_{t-1})$   is (9)
             $E_t = |L(\mathfrak{g}_t) - L(\mathfrak{g}_{t-1})|$
             Increment
          END UNTIL
          $(P, \mathfrak{g}) \leftarrow (P_t, \mathfrak{g}_t)$;
```

Appendix 2
The FCM Algorithm used to initialize EM [2]

```
Store     Unlabeled Object Data $X \in \Re^p$

Pick      ● number of clusters: $1 < c < n$. Rule of thumb: $c \leqslant \sqrt{n}$
          ● max. # of iterations: $T = 300$
          ● weighting exponent: $m = 2$
          ● norm for clustering criterion $J_2$: $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$
          ● termination measure: $E_t = \underbrace{\max}_{\substack{1 \leqslant i \leqslant c \\ 1 \leqslant s \leqslant p}} \{|v_{si,t} - v_{si,t-1}|\}$
          ● termination criterion: $\varepsilon = 0.01$

Get       ● initial cluster centers: $\mathbf{V}_0 \in \Re^{cp}$: 3 methods
          $m_j = \underbrace{\min}_{k} \{x_{jk}\}: j = 1, 2, \ldots, p$
          $M_j = \underbrace{\max}_{k} \{x_{jk}\}: j = 1, 2, \ldots, p$
          I1.   $\mathbf{v}_{i,0} = \mathbf{m} + \left(\frac{i-1}{c-1}\right)(\mathbf{M} - \mathbf{m}))$; $i = 1, 2, \ldots, c$
          I2; I3.   Randomly draw two sets of $\{\mathbf{v}_{i0}\}$'s from $hb(\mathbf{m.M})$

Iterate   $0 \leftarrow t$
          DO UNTIL ($t = T$ or $E_t \leqslant \varepsilon$):
          $u_{ik,t} = \left[\sum_{j=1}^{c} (\|\mathbf{x}_k - \mathbf{v}_{i,t-1}\|_A / \|\mathbf{x}_k - \mathbf{v}_{j,t-1}\|_A)^{\frac{2}{m-1}}\right]^{-1} \forall i,k$
          $\mathbf{v}_{i,t} = \sum_{k=1}^{n} (u_{ik,t})^m \mathbf{x}_k \Big/ \sum_{k=1}^{n} (u_{ik,t})^m$
          $E_t = \|\mathbf{V}_t - \mathbf{V}_{t-1}\|_\infty$
             Increment t
          END UNTIL
          $P_0 \leftarrow U$
```

References

1. **Titterington, D.; Smith, A.; Makov, U.:** Statistical Analysis of Finite Mixture Distributions, Wiley, NY, 1985
2. **Bezdek, J.C.:** Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY, 1981
3. **Krishnapuram, R.; Keller, J.:** A Possibilistic Approach to Clustering, *IEEE Trans. Fuzzy Systems*, 1(2), 98–110, 1993
4. **Jain, A.; Dubes, R.:** Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, 1988
5. **Morrison, M.; Attikouzel, Y.:** An introduction to the segmentation of Magnetic Resonance images, *Aust. Comp. Jo.*, 26(3), 90–98, 1994
6. **Bezdek, J.C.; Hall, L.O.; Clarke, L.P.:** Review of MR image segmentation techniques using pattern recognition, Med. Physics, 20, 1033–1047, 1993
7. **Dubes, R.; Jain, A.:** Clustering techniques: the user's dilemma, Patt. Recognition, 8, 247–260, 1977
8. **Bock, H.H.:** On some significance tests in cluster analysis, *Jo. Classification*, 2, 77–108, 1985
9. **Milligan, G.; Cooper, M.C.:** An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50, 159–179, 1985
10. **Soromenho, G.:** Comparing approaches for testing the number of components in a finite mixture model, Comp. Statistics, 65–78, 1994
11. **Cutler, A.; Windham, M.P.:** Information-based validity functionals for mixture analysis, in Proc. 1st US/Japan Conf. on the Frontiers of Statistical Modeling, ed. H. Bozdogan, Kluwer, Amsterdam, 149–170, 1994
12. **Celeux, G.; Soromenho, G.:** An entropy criterion for assessing the number of clusters in a mixture model, Jo. Classification, 13(2), 195–212, 1996
13. **Bezdek, J.C.; Windham, M.; Ehrlich, R.:** Statistical Parameters of Fuzzy Cluster Validity Functionals , Int. Jo. Comp. and Inf. Sci., 9(4), 1980, 232–336, 1980
14. **Windham, M.P.:** Cluster validity for the fuzzy c-means clustering algorithm, IEEE Trans. PAMI, 4(4),357–363, 1982
15. **Pal, N.R.; Bezdek, J.C.:** On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Systems, 3(3), 370–376, 1995
16. **Wolfe, J.H.:** Pattern clustering by multivariate mixture analysis, Multivariate Behavioral Research, 5, 329–350, 1970
17. **Bezdek, J.C.; Hathaway, R.J.; Howard, R.E.; Wilson, C.A.:** Coordinate Descent and Clustering, Control and Cybernetics, 15(2), 195–203, 1996
18. **Bezdek, J.C.; Hathaway, R.J.; Howard, R.E.; Wilson, C.A.; Windham, M.P.:** Local Convergence Analysis of a Grouped Variable Version of Coordinate Descent, J. Optimization Theory and Appl., 54(3), 471–477, 1987
19. **Hathaway, R.J.; Redner, R.; Bezdek, J.C.:** Estimating the Parameters of Mixture Models with Modal Estimators, Comm. in Stat. (A), 16(9), 1987, 2639–2660, 1987
20. **Hathaway, R.J.; Bezdek, J.C.:** Grouped Coordinate Minimization using Newton's Method for Inexact Minimization in One Vector Coordinate, J. Optimization Theory and Appl., 71(3). 71(3), 503–516, 1991
21. **Duda, R.; Hart, P.:** Pattern Classification and Scene Analysis, Wiley Interscience, NY, 1973
22. **McLachlan, G.J.; Basford, K.E.:** Mixture Models: Inference and Applications, Marcel Dekker, Inc., New York and Basel, 1988
23. **Wallace, C.S.; Boulton, D.M.:** An information measure for classification, Computer Journal, 11(2), 185–194, 1968
24. **Akaike, H.:** A new look at the statistical model identification. IEEE Trans. Automatic Control, 19(6), 716–723, 1974
25. **Bozdogan, H.:** Choosing the number of component clusters in the Mixture-Model Using a New Information Complexity Criterion of the Inverse-Fisher Information Matrix, Studies in Classification, Data Analysis, and Knowledge Organization, ed. O. Opitz, B. Lausen, and R. Klar, Springer-Verlag, Heidelberg, 40–54, 1993
26. **Rissanen, J.:** Modeling By Shortest Data Description, Automatica, 14, 465–471, 1978
27. **Liang, Z.; Jaszczak, R.J.; Coleman. R.E.:** Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing, IEEE Trans. Nuclear Science, 39(4), 1126–1133, 1992
28. **Banfield, J.; Raftery, A.E.:** Model-based gaussian and non-gaussian clustering, Biometrics, 49, 803–821, 1993
29. **Bozdogan, H.:** Mixture-Model Cluster Analysis Using Model Selection Criteria and A New Informational Measure of Complexity, Proc. First US/Japan Conference on the Frontiers of Statistical Modeling: An Information Approach, Bozdogan, H. (ed.), 69–113, 1994

30. **Roubens, M.:** Fuzzy Clustering Algorithms and Their Cluster Validity, Eur. Jo. Op. Res., 10, 294–301, 1982
31. **Xie, X.L.; Beni, G.A.:** Validity Measure for Fuzzy Clustering, IEEE Trans. PAMI, 3(8), 841–846, 1991
32. **Windham, M.P.; Cutler, A.:** Information ratios for validating mixture analysis. JASA, 87(420), 1188–1192, 1992
33. **Hubert, L.J.; Arabie, P.:** Comparing partitions, *J. Classification*, 2, 193–218, 1985
34. **Davies, D.L.; Bouldin, D.W.:** A cluster separation measure, IEEE Trans. PAMI, 224–227, 1979
35. **Dunn, J.C.:** A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Jo. Cybernetics, 3(3), 32–57, 1974
36. **Bezdek, J.C.; Pal, N.R.:** Cluster validation with generalized Dunn's indices, Proc. 1995 2nd NZ Int'l. two-stream conference on ANNES, ed. N. Kasabov and G. Coghill, IEEE Press, Piscataway, NJ, 190–193, 1995
37. **Clark, M.; Hall, L.; Goldgof, D.; Clarke, L.; Velthuizen, R.; Silbiger, M.:** MRI segmentation using fuzzy clustering techniques: integrating knowledge, IEEE Engineering in Medicine and Biology Magazine, 13(5), 730–742, 1994
38. **Di Gesu, V.; De La Paz, R.; Hanson, W.A.; Berstein, R.:** Clustering algorithms for MRI, in Lecture notes for medical informatics, K.P. Adlassing, B. Grabner, S. Bengtsson and R. Hansen, eds., Springer, 534–539, 1991
39. **Jain, A.K.; Flynn, P.J.:** Image segmentation using clustering, in Advances in Image Understanding, eds. K. Bowyer and N. Ahuja, IEEE Computer Society Press, Los Alamitos, CA, 65–82, 1996
40. **McKenzie, P.; Alder, M.:** Initializing the EM algorithm for use in Gaussian mixture modeling, in Pattern Recognition in Practice IV; Multiple Paradigms, Comparative Studies and Hybrid Systems, ed. E.S. Gelsema and L. N. Kanal, Elsevier, NY, 91–105, 1994
41. **Bezdek, J.C.; Dunn, J.C.:** Optimal fuzzy partitions: a heuristic for estimating the parameters in a mixture of normal distributions, IEEE Trans. Comp., 24(8), 835–838, 1975
42. **Bezdek, J.C.; Hathaway, R.J.; Huggins, V.:** Parametric Estimation for Normal Mixtures, Patt. Recog. Letters, 3, 1985, 79–84, 1985
43. **Davenport, J.; Bezdek, J.C.; Hathaway, R.J.:** Parameter Estimation for Finite Mixture Distributions, Int'l. Jo. Comp. & Math. with Applications, 15(10), 819–828, 1988