An Objective Approach to Cluster Validation

Mohamed Bouguessa [a], Shengrui Wang, [a,*], Haojun Sun [b]

[a] *Département d'Informatique, Faculté des sciences, Université de Sherbrooke, Sherbrooke, Québec, J1K 2R1, Canada*

[b] *College of Mathematics and Computer Sciences, Hebei University, Baoding, 071002, China*

Abstract

Cluster validation is a major issue in cluster analysis. Many existing validity indices do not perform well when clusters overlap or there is significant variation in their covariance structure. The contribution of this paper is twofold. First, we propose a new validity index for fuzzy clustering. Second, we present a new approach for the objective evaluation of validity indices and clustering algorithms. Our validity index makes use of the covariance structure of clusters, while the evaluation approach utilizes a new concept of overlap rate that gives a formal measure of the difficulty of distinguishing between overlapping clusters. We have carried out experimental studies using data sets containing clusters of different shapes and densities and various overlap rates, in order to show how validity indices behave when clusters become less and less separable. Finally, the effectiveness of the new validity index is also demonstrated on a number of real-life data sets.

Key words: Fuzzy clustering, Validity index, Overlapping clusters, Overlap rate, Truthed data set.

1 Introduction

The aim of clustering techniques is to partition a given set of data or objects into groups such that elements drawn from the same group are as similar to each other as possible, while those assigned to different groups are dissimilar. There are many methods and algorithms for clustering based on crisp (Duda

* Corresponding author.
Email addresses: m.bouguessa@usherbrooke.ca (Mohamed Bouguessa), s.wang@usherbrooke.ca (Shengrui Wang,), haojun.sun@mail.hbu.edu.cn (Haojun Sun).

et al., 2001), fuzzy (Bezdek, 1981; Hoppner et al., 1999), probabilistic (Titterington et al., 1985), and possibilistic approaches (Krishnapuram and Keller, 1993). The clustering process considered in this paper is fuzzy clustering, in particular the Fuzzy Maximum Likelihood Estimation algorithm (FMLE) (Gath and Geva, 1989), because of its efficiency in dealing with variation in cluster shapes and densities.

In practical applications of clustering algorithms, several problems must be solved, including determination of the number of clusters and evaluation of the quality of the partitions. To address these problems, many functions called cluster validity indices have been proposed in the literature (Pal and Biswas, 1997; Rezae et al., 1998; Geva et al., 2000; Maulik and Bandyopadhyay, 2002; Sun et al., 2004). A validity index provides an objective measurement of a clustering result and its optimal value is often used to indicate the best possible choice for the values of parameters in the clustering algorithm (e.g., the number of clusters).

The performance of validity indices is often assessed based on how well they deal with situations in which clusters overlap or there are significant variations in cluster shape, density and orientation. Such cases arise often in practical applications. For instance, in color image segmentation, a "region" may correspond to a cluster of pixels with similar spectral properties (a cluster of similar pixels may also encompass several regions). Two similar regions correspond to two overlapping clusters. Whether and how such overlapped clusters are separated has a direct impact on the results of segmentation. Due to the lack of understanding of component overlapping in a mixture, most researchers evaluate the results of validity indices using *ad hoc* data sets. There is a need for a more systematic approach to the comparison of validity indices, based on formal data distribution model. This need motivated our efforts to present a new method for generating test data sets with different degrees of overlap between clusters. The generated data sets allow a precise account of the dependence of an index's performance on the separation between clusters.

It is important to note that the Gaussian mixture is a fundamental hypothesis that many partition-based clustering algorithms make regarding the data distribution model. Aitnouri et al. (2000), explicitly defined the concept of overlap rate in the 1-dimensional case and developed algorithms for generating data sets with overlapped clusters. In (Sun and Wang, 2003), we established a general theory of the degree of overlap between a pair of components of a Gaussian mixture for the multi-dimensional case. The theory in (Sun and Wang, 2003) is important in that it provides a physical measure of the complexity of a data set and lays down a foundation for controlling the degree of overlap between clusters as a function of the parameter values of each component. Consequently, it allows the generation of truthed data sets of different levels of difficulty that can be used to compare the performance of clustering

algorithms and validity indices. By "truthed data set", we mean that the generated data set contains distinguishable clusters according to the measure of overlap rate (OLR) presented in Section 4.

This paper is organized as follows. In Section 2, we introduce the fuzzy c-means (FCM) and FMLE algorithms. We also discuss an implementation strategy for determining the number of clusters using these algorithms. In Section 3, we present our new validity index, following a description of several major existing validity indices. In Section 4, the theory regarding the overlap rate is introduced. We show how this theory can be used to generate truthed data sets as well as to measure the overlap rate of a given data set. In Section 5, we present several generated Gaussian-mixture data sets with different overlap rates and report experimental studies comparing the performance of the validity indices. The performance of the proposed index is also tested on a number of real-world data sets. Section 6 presents our conclusions.

2 Clustering Algorithms

In this section, we briefly introduce the FCM algorithm and describe the clustering process based on the FMLE algorithm. We also provide some arguments justifying our choice of the FMLE algorithm for this work.

2.1 FCM algorithm

The FCM algorithm, introduced by Dunn (1973) and generalized by Bezdek (1981), is the fuzzy clustering algorithm most widely used in practice. It is based on an iterative optimization of a fuzzy objective function:

$$Minimize\ J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}^m \|x_k - v_i\|^2, \tag{1}$$

where n is the total number of data vectors in a given data set and c is the number of clusters; $X = \{x_1, x_2, \cdots, x_n\} \subset \Re^d$ and $V = \{v_1, v_2, \cdots, v_c\} \subset \Re^d$ are the feature data and cluster centers; and $U = (u_{ki})_{n\times c}$ is a fuzzy partition matrix composed of the membership of each feature vector x_k in each cluster i, where u_{ki} satisfies $\sum_{i=1}^{c} u_{ki} = 1$ for $k = \{1, 2, \cdots, n\}$ and $u_{ki} \geq 0$ for all $i = \{1, 2, \cdots, c\}$ and $k = \{1, 2, \cdots, n\}$. The exponent $m > 1$ in Eq.(1) is a parameter, usually called a *fuzzifier*.

The distance measure used is the Euclidean distance between a datum and a prototype (cluster center). Consequently, the resulting fuzzy c-means algo-

rithm recognizes only circular and spherical clusters. By replacing the Euclidean distance in FCM by another metric derived from a positive, definite, symmetric matrix, ellipsoidal clusters can also be recognized. FMLE is such an efficient algorithm, which performs well in situations where the variability of cluster shapes and densities is great.

2.2 FMLE algorithm

Gath and Geva (1989) proposed the Fuzzy Maximum Likelihood Estimation algorithm (FMLE). FMLE extends FCM by the computation of a fuzzy covariance matrix for each cluster, thereby allowing the recognition of elliptical clusters. For this purpose an exponential distance measure is defined based on maximum likelihood estimation (Bezdek and Dunn, 1975).

The target data sets of the FMLE algorithm are those that follow the distribution of a general mixture of Gaussians. This is the main reason why the algorithm has been selected for this study. However, it is helpful to initialize the FMLE algorithm with the membership matrix generated by the FCM algorithm in order to reduce the number of iteration steps.

2.3 Determination of the number of clusters

The algorithm below follows from the general model selection approach to searching for the optimal c-partition of a data set X, given the minimal and maximal number of clusters C_{min} and C_{max} (C_{min} and C_{max} are predefined). $V_d(c)$ in the algorithm is the value of the validity function to be optimized.

Algorithm 1: FMLE-based model selection algorithm

(1) Choose $C_{\min}$ and $C_{\max}$.
(2) For $c = C_{\min}$ to $C_{\max}$:
  (a) Apply the FCM algorithm as an initialization of the FMLE algorithm.
  (b) Apply the FMLE algorithm.
  (c) Compute a validity value $V_d(c)$.
(3) Compute c_f such that $V_d(c_f)$ is optimal.
(4) End.

3 Cluster Validity

Validity indices are extremely important for automatically determining the number of clusters. Various validity indices have been proposed in the past. In the following subsection, we present a number of existing indices and discuss the main ideas underlying them. A new validity index is then introduced.

3.1 Validity indices for fuzzy clustering

In general, validity indices can be grouped into three main categories. Those in the first category use only the property of fuzzy membership degree to evaluate a partition. Those in the second group combine the property of membership degree and the geometric structure of the data set, while those in the third make use of the concepts of hypervolume and density. Some of the indices most frequently referred to in the literature are described below.

The partition coefficient V_{PC} (to be maximized) and the partition entropy V_{PE} (to be minimized), described by Bezdek et al. (1999), are examples of the first category. Both indices are computed using only the elements of the membership matrix. Their main disadvantage is their lack of direct connection to the geometrical properties of the data and their monotonic dependency on the number of clusters (Hoppner et al., 1999). For our experiments we have chosen V_{PC}.

A well-known index from the second category of validity indices is the Fakuyama and Sugeno validity V_{FS} (Pal and Bezdek, 1995), which measures the discrepancy between compactness and separation of clusters. The number of clusters that minimizes V_{FS} is taken as the optimal number of clusters.

Xie and Beni (1991) proposed another well-known validity index, V_{XIE}, which measures overall average compactness against separation of the c-partition. The main disadvantage of this index is that it tends to decrease monotonically when c is very large (Xie and Beni, 1991). Smaller V_{XIE} means a more compact and well-separated c-partition.

Zahid et al. (1999) proposed the validity index V_{ZLE}, based on a combination of two functions, each of which is given by a fuzzy separation-compactness ratio. The first function calculates this ratio by considering the geometrical properties and membership degrees of the data. The second evaluates it using only the property of membership degree. The maximum of V_{ZLE}, as a function of the number of clusters c, is sought for a well-defined c-partition.

Geva et al. (2000) introduce a number of scattering criteria derived from the

scatter matrices used in discriminant analysis (Fukunaga, 1990). For purposes of comparison, we have chosen the normalized invariant criterion V_{N_INV}. This index measures the trace of the product matrix between the inverse of the within-cluster scatter matrix and the between-cluster scatter matrix normalized by the number of clusters c^2, as presented in Table 1. The cluster number that maximizes V_{N_INV} is considered to be the optimal value for the number of clusters present in the data.

Xie et al. (2002) proposed an index V_{XRZ} based on the separation-compactness ratio. The maximum of V_{XRZ}, as a function of the number of clusters c, is sought for a well-defined c-partition. Sun et al. (2004) proposed an index V_{WSJ} that measures the separation between clusters and the cohesion within clusters. This index is based on a linear combination of the average within-cluster scattering (inversely related to compactness) and between-cluster distance (separation). A cluster number which minimizes V_{WSJ} corresponds to the best clustering.

The last indices covered here, from the third category, are those of Gath and Geva (1989), who introduce three validity indices based on the concepts of hypervolume and density. For our experiments we have chosen two of these. The first one is the fuzzy hypervolume, V_{FH}, which considers the sum of all cluster sizes. A good partition should yield a low fuzzy hypervolume. The second one is the average partition density, V_{APD}, in which the fuzzy density is calculated for each cluster and then averaged over all clusters. The cluster number that maximizes V_{APD} is considered to be the optimal value for the number of clusters present in the data. Table 1 lists all of these validity indices.

3.2 A new validity index

In this subsection, we propose a new validity index based on the general principle of optimizing a combined function of compactness and separation. This principle has been followed by most of researchers in their efforts to develop validity indices. Similar to several recently proposed indices, the new index presented here makes use of existing or slightly modified definitions for the concepts of separation and compactness. We believe that its judicious combination of all of the relevant factors has made the new index particularly efficient in various situations, as shown by the experiments reported in Section 5. Although there is no formal way to prove the efficiency of a validity index (which is our reason for proposing a new method for generating test data sets in this paper), we will try to provide some insights that justify the proposed index.

Table 1
Nine validity indices for the Fuzzy Clustering.

Validity index	Functional description
$V_{PC}(U)$	$\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}^2$
$V_{FS}(U,V,X)$	$\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}^m\,(\|x_k-v_i\|^2-\|v_i-\bar{v}\|^2)\,;\bar{v}$ *is the overall mean of the data set*
$V_{XIE}(U,V,X)$	$(\sum_{k=1}^{n}\sum_{i=1}^{c}[(u_{ki}^m)\,\|x_k-v_i\|^2])/\,(n\,[\min_{i\neq j}(\|v_i-v_j\|^2)])$
$V_{ZLE}(U,V,X)$	$SC_1(U,V,X)+SC_2(U)$
	$SC_1(U,V,X)=\left(\frac{1}{C}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2\right)\Big/\left(\sum_{i=1}^{c}\left[\frac{(\sum_{k=1}^{n}(u_{ki})^m\|x_k-v_i\|^2)}{\sum_{k=1}^{n}u_{ki}}\right]\right)$
	$SC_2(U)=\left(\sum_{i=1}^{c-1}\sum_{r=1}^{c-i}\left(\frac{\sum_{k=1}^{n}[min(u_{ki},u_{kj})]^2}{\sum_{k=1,j=r+i}^{n}min(u_{ki},u_{kj})}\right)\right)\Big/\left(\frac{\sum_{k=1}^{n}[\max_{1\le i\le c}u_{ki}]^2}{\sum_{k=1}^{n}\max_{1\le i\le c}u_{ki}}\right)$
$V_{N\_INV}(U,V,X)$	$trace(S_W^{-1}.S_B)/c^2$
	$S_W=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ki}(x_k-v_i)(x_k-v_i)^T$
	$S_B=\sum_{i=1}^{c}(\sum_{k=1}^{n}u_{ki})(v_i-\overline{v})(v_i-\overline{v})^T$
$V_{XRZ}(U,V,X)$	$\frac{nc}{c}.SP(V).CP(U,V,X)$
	$SP(V)=\frac{1}{c^2}\left(\sum_{i=1}^{c}\min_{1\le j\le c,i\neq j}\|v_i-v_j\|\right)^2$
	$CP(U,V,X)=nc\left(\sum_{i=1}^{nc}\left[\left(\sum_{x_k\in p_i,x_k\neq v_i}(u_{ki})^2\|x_k-v_i\|^2\right)\Big/\sum_{x_k\in p_i,x_k\neq v_i}(u_{ki})^2\right]\right)^{-1}$
	$where\ \ nc=\|C\text{`}\|$
	$C\text{`}=\{C_{pi}\|C_{pi}\in C\ ;\ C_{pi}$ *is not singleton,* $i=(1,\ldots,nc),\ C=(C_1,\ldots,C_c)\}$
$V_{WSJ}(U,V,X)$	$Scat(c)+(Sep(c)/\ Sep(C_{max}))$ $\ ;C_{max}$ *is the maximum number of clusters*
	$Scat(c)=(\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|)/\,(\|\sigma(X)\|)$
	$\sigma(X)=\left\{\sigma(X)^1,\ldots,\sigma(X)^d\right\}^T,\ \sigma(X)^p=\frac{1}{n}\sum_{k=1}^{n}(x_k^p-\bar{x}^p)^2,\ \bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$
	$\sigma(v_i)=\left\{\sigma(v_i)^1,\ldots,\sigma(v_i)^d\right\}^T,\sigma(v_i)^p=\frac{1}{n}\sum_{k=1}^{n}u_{ki}(x_k^p-v_i^p)^2,p=(1,\ldots,d)$
	$Sep(c)=\frac{D_{max}^2}{D_{min}^2}\sum_{i=1}^{c}\left(\sum_{j=1}^{c}\|v_i-v_j\|^2\right)^{-1}$
	$D_{min}=\min_{i\neq j}\|v_i-v_j\|\ and\ D_{max}=\max_{i,j}\|v_i-v_j\|$
$V_{FH}(\Sigma)$	$\sum_{i=1}^{c}[\sqrt{det(\Sigma_i)}],\ \ \Sigma_i=\sum_{k=1}^{n}u_{ki}\,(x_k-v_i)\,(x_k-v_i)^T\Big/\sum_{k=1}^{n}u_{ki}$
$V_{APD}(U,\Sigma)$	$\frac{1}{c}\sum_{i=1}^{c}\frac{S_i}{\sqrt{det(\Sigma_i)}},\ \ S_i=\sum_{k\in w_i}u_{ki},\quad w_i=\{k\in N_{\le n}:[(x_k-v_i)^T\Sigma_i^{-1}(x_k-v_i)]<1\}$
	$\Sigma_i=\sum_{k=1}^{n}u_{ki}\,(x_k-v_i)\,(x_k-v_i)^T\Big/\sum_{k=1}^{n}u_{ki}$

The new validity index is defined as follows:

$$V_{SC}(U, V, X) = Sep(c) \Big/ Comp(c) \tag{2}$$

In this formula, $Sep(c)$ is the fuzzy separation of fuzzy clusters given by:

$$Sep(c) = trace(S_B) \tag{3}$$

where S_B is the between-cluster fuzzy scatter matrix, defined as:

$$S_B = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ki}^m (v_i - \bar{v})(v_i - \bar{v})^T \tag{4}$$

Here, S_B is defined in the same way as in the index V_{N_INV} (listed in Table 1). A large value of $Sep(c)$ indicates that the fuzzy c-partition is characterized by well-separated fuzzy clusters.

$Comp(c)$, the total compactness of the fuzzy c-partition, is given by:

$$Comp(c) = \sum_{i=1}^{c} trace(\Sigma_i) \tag{5}$$

where Σ_i is the fuzzy covariance matrix, defined as:

$$\Sigma_i = \Big(\sum_{k=1}^{n} u_{ki}^m \, (x_k - v_i) \, (x_k - v_i)^T \Big) \Big/ \sum_{k=1}^{n} u_{ki}^m \tag{6}$$

In the above definition, the individual compactness of each cluster is measured by the trace of its covariance matrix (see Eq.(5)). The numerator on the right side of Eq.(6) is the conventional compactness matrix (Pal and Bezdek, 1995; Duda et al., 2001). The use of the covariance matrix, which is obtained by normalization of this compactness matrix by $\sum_{k=1}^{n} u_{ki}^m$, makes $Comp(c)$ more sensitive to the variation of the covariance structure of each cluster and avoids the monotony problem, as remarked by Geva et al. (2000). Because of the linearity of the trace operation, $Comp(c)$ can be written as $Comp(c) = trace(S_{W'})$ if we define our within-cluster scatter matrix by $S_{W'} = \sum_{i=1}^{c} \Sigma_i$. A small value of $Comp(c)$ indicates a compact partition. So a compact and separate c-partition corresponds to a large value of V_{SC}. In other words, V_{SC} is to be maximized.

Apart from their respective definitions, the way in which compactness and separation are combined also plays a critical role in the performance of the new

index. In fact, two formula based on the well-known Fisher separability criterion (Fukunaga, 1990) are $J_1 = trace(S_B)/trace(S_{W'})$ and $J_2 = trace(S_{W'}^{-1}S_B)$. We have adopted J_1 for our index (rather than J_2 as is the case of V_{N_INV}), for two reasons. One is computational simplicity since J_1 is obviously easier to compute than J_2. The main reason for our choice, however, is that the ratio approach measures separation and compactness independently before combining them. This is important because in J_2, certain parameters such as individual cluster orientations may interact with each other and thus, when they change from one data set to another, variations in J_2 may result even though the variation of these parameters does not have any significant impact on separation and compactness. A theoretical explanation of this phenomenon is an open question, although it is possible to provide a formal analysis of some simple cases. For the sake of this paper's focus, we only illustrate the phenomenon by examples.

We generate a group of ten Gaussian mixture data sets $(X_1, \ldots, X_{10})$. Each data set contains three well-separated elliptical clusters in $\Re^3$. The first data set X_1 is illustrated in Fig. 1. The data sets $X_2, \ldots, X_{10}$ are, in turn, obtained by pivoting each of the clusters 2 and 3 by 10 degrees around the x-axis. We assume that cluster centers are perfectly localized in this experiment (they are known by virtue of the data generation procedure). The values of u_{ki} used by S_B and $S_{W'}$ are obtained by the same formula in the FMLE algorithm. Fig. 2 illustrates the variation of J_1 and J_2 as a function of $X_1, \ldots, X_{10}$. Clearly the function J_1 remains almost constant while J_2 is unstable for the group of data sets.

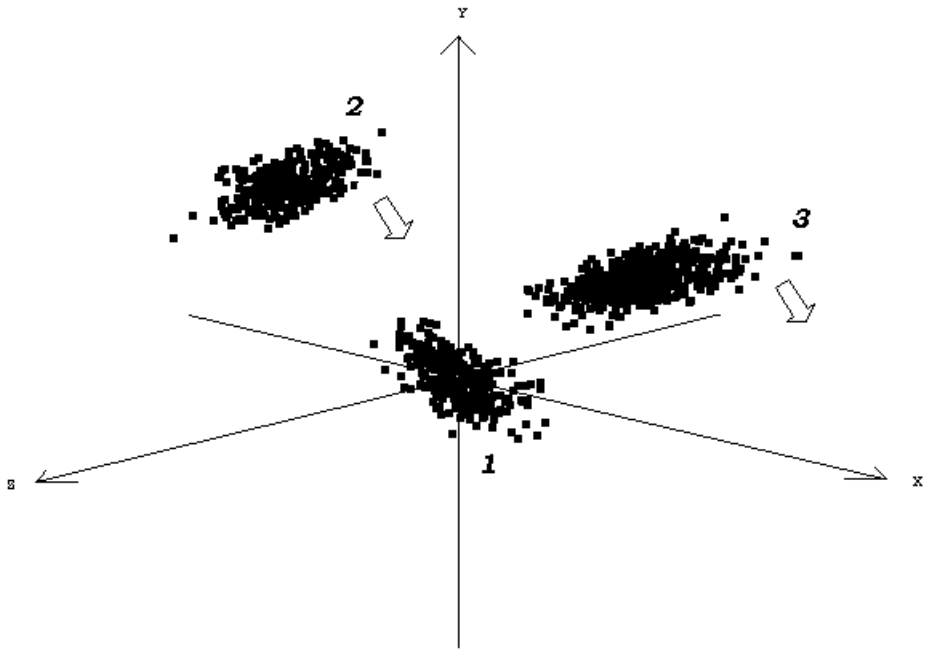

Fig. 1. Data set X_1; the other sets are obtained by pivoting cluster 2 and 3

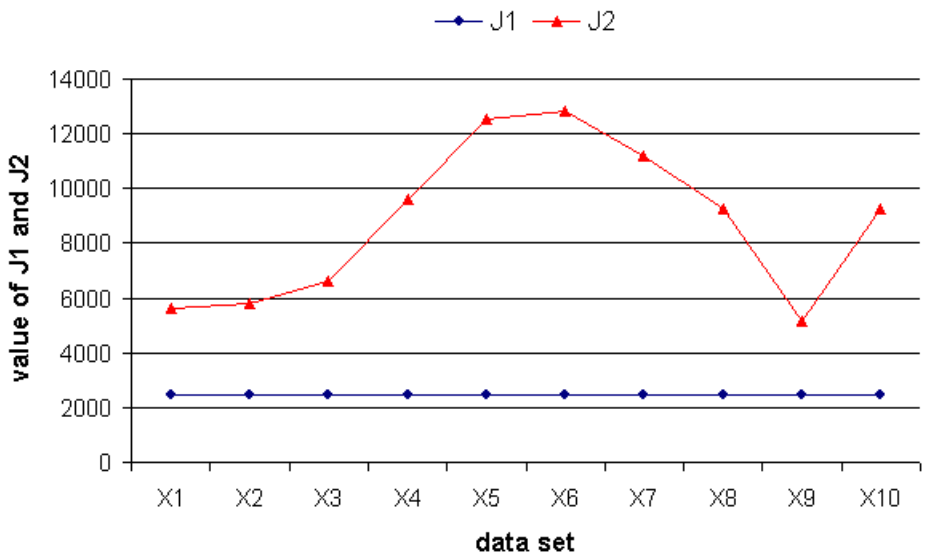

Fig. 2. Variation of J_1 and J_2

This example suggests that, compared to J_2, J_1 is more directly dependent on the separation and compactness values of individual clusters while J_2 can be more easily affected by parameters whose value changes may not have significant impact on the two basic measures for cluster analysis. Here, we believe that the formula based on the definition of J_1 should result in an index more resistant to the variation of some factors (e.g. orientation) relating to the distribution of clusters than an index based on the definition of J_2. The suitability of the proposed index, V_{SC}, was tested on number of difficult data sets in Section 5. The "difficulty" here means overlap between clusters and will be formally introduced in the following section.

4 Measuring Cluster Overlap

4.1 Theoretical framework

There cannot be a mathematical proof that one index is better than another. The only way to demonstrate the performance of a validity index is to test it on concrete data sets. That is why it is extremely important to have truthed data sets (i.e., sets for which the number of clusters and the values of the cluster parameters are known) and a formal measure of data complexity (how well clusters are separated from each other). In practice, all of the existing indices work well on data sets containing only well-separated spherical clusters, but many of them fail if the data set contains overlapping clusters. The ability to deal with overlapping clusters is considered to be one of the main advantages distinguishing one index from another. Despite this importance, almost

all of the reported work is based on an intuitive account of the overlapping phenomenon.

Given a data set whose distribution corresponds to a mixture of Gaussians, the degree of overlap between components affects the number of clusters perceived by a human operator or detected by a validity index. In other words, there may be a significant difference between intuitively defined clusters and the true clusters corresponding to the components in the mixture. The component overlapping phenomenon is illustrated in Fig. 3.

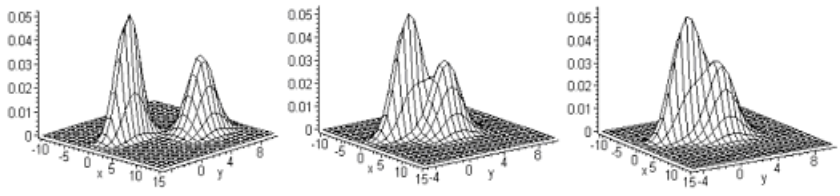

Fig. 3. Two components that are well separated, partially overlapping and strongly overlapping.

In general, one expects a clustering algorithm and validity index to be able to distinguish between partially overlapping clusters. In order to precisely measure the overlap rate as well as to generate truthed data sets containing overlapping clusters with prescribed overlap rates, we have proposed a theoretical framework for the Gaussian mixture model (Sun and Wang, 2003). The theory is based on a novel ridge curve concept and establishes a series of theorems characterizing the overlap phenomenon as a function of parameters of the mixture.

In its simplest form, the *pdf* (probability density function) of a mixture of two Gaussian components in the d-dimensional space is given by:

$$P(x) = \sum_{i=1,2} \alpha_i G_i(x, \mu_i, \Sigma_i), \quad x = (x_1, x_2, \ldots, x_d) \tag{7}$$

with the restriction $\alpha_i > 0$, where α_i is the mixing coefficient and $\sum_{i=1}^{2} \alpha_i = 1$. Note that μ_i and Σ_i denote, respectively, the mean and the covariance matrix for the ith distribution G_i.
G_i is the ith component, given by:

$$G_i(x, \mu_i, \Sigma_i) \;=\; (2\pi)^{-\frac{d}{2}} \; |\Sigma_i|^{-\frac{1}{2}} \; exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)) \tag{8}$$

To study the overlap phenomenon, we consider the case of two components in a mixture. For a mixture with multiple components, one would need to measure the degree of overlap between each pair of components.

The overlap rate (OLR) implements the following principle: 1) OLR decreases ($\rightarrow 0$) as the two components become more widely separated; 2) OLR increases ($\rightarrow 1$) as the two components become more strongly overlapped.

Definition 1 *The overlap rate between two components of the mixture, G_1 and G_2, is determined by the ratio between the values of the peak and saddle of the pdf.*

$$OLR(G_1, G_2) = \begin{cases} 1 & pdf\ has\ one\ peak \\ \frac{P(x_{saddle})}{P(x_{lowerpeak})} & pdf\ has\ two\ peaks \end{cases} \tag{9}$$

OLR provides a formal measure of the difficulty in distinguishing between two components. It depends not only on the distance between the two component means but also on the shape, orientation and density of each of the components. Computation of OLR is simplified thanks to the concept of the ridge curve. The ridge curve in the 2-D case is defined as:

$$A_{x_1} B_{x_2} - A_{x_2} B_{x_1} = 0 \tag{10}$$

where

$$A_{x_1} = \frac{\partial}{\partial x_1}(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)) \tag{11}$$

$$B_{x_1} = \frac{\partial}{\partial x_1}(-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)) \tag{12}$$

In the general multi-dimensional case (supposing that d is the number of dimensions), the ridge curve is defined by a system of d-1 equations of the same type as Eq. (10). The ridge curve concept makes it possible to search for the peak and saddle points along a curve rather than in a region of the original space. The following theorems give the properties of the ridge curve.

Theorem 1 *The ridge curve is a hyperbola or a line.*

Theorem 2 *The means of the two components and the stationary points (peak points and saddle points) of the pdf are on the ridge curve.*

Theorem 3 *The stationary points of the pdf fall on the segment between the two means of the components of the ridge curve.*

4.2 Algorithm for calculating OLR

Using the previous theorems, we can calculate *OLR* based on a linear search. The main idea of the algorithm is to search the curve defined in Eq.(10) between the means of the two components. A local maximum point is a peak of the *pdf*, and the minimum point is a saddle point. Algorithm 2 below computes the overlap rate of two mixture components, defined in Eq.(7).

Algorithm 2: OLR between two components

(1) Compute the parameters of two distributions: the means (μ_1, μ_2), the covariance matrices (Σ_1, Σ_2) and the prior probability of each cluster (α_1, α_2).
(2) Compute the ridge curve using Eq.(10).
(3) Move from μ_1 to μ_2 on the ridge curve, finding the maximum and minimum points of $P(x)$.
(4) Compute *OLR* for the two components using Eq.(9).
(5) End.

On the other hand, since *OLR* is a function of the parameters of the Gaussian mixture model, varying these parameters (including the mixing coefficient and the distance between the two means and the two covariance matrices) will affect the theoretical value of *OLR*. Many parameters can be modified to generate a variety of truthed data sets containing overlapping clusters with different covariance structures. In order to illustrate this phenomenon, we generate a number of truthed data sets in Section 5.

To close this section, it is worth pointing out that the (Gaussian) classification error might have been used here as a measure of overlap between two clusters. We prefer *OLR* because it is computationally more efficient and corresponds better to a geometrical interpretation of the overlapping phenomenon. The latter point is illustrated by the following example. Let us consider two Gaussian mixtures $G(x, y) = 0.5G_1(x, y) + 0.5G_2(x, y)$ and $\grave{G}(x, y) = 0.7G_1(x, y) + 0.3G_2(x, y)$, with $G_1(x, y) = \frac{1}{2\pi} exp(-\frac{1}{2}(x^2 + y^2))$ and $G_2(x, y) = \frac{1}{2\pi} exp(-\frac{1}{2}((x - 2.7)^2 + y^2))$. *OLR* in $G(x, y)$ is 0.781 and in $\grave{G}(x, y)$ is 1, while the classification error rate between the two components in $G(x, y)$ is 0.085 and in $\grave{G}(x, y)$ is 0.0786. These results mean that in the graphic plot of the two mixtures, there is a saddle point in $G(x, y)$ lower than either of the peaks near the component centers (theoretically, the peak is not at the center of a component); whereas in $\grave{G}(x, y)$, there is only one peak. These features are reflected in the respective *OLR* values, but not in the classification errors. Quite the opposite: the classification error gives a smaller value for $\grave{G}$ than for G. Thus, this example shows that the proposed *OLR* and the classification error may evolve in opposite directions when the value of some parameter

(the mixing coefficient in this case) is changed. A more extensive comparison between the two measures is beyond the scope of this paper.

5 Objective Evaluation of Validity Indices

In this section, we present a comparative evaluation of all the validity indices discussed in Section 3, to illustrate their effectiveness in finding an optimal cluster scheme in the presence of overlapping clusters that differ in density, shape and orientation. We first describe the different data sets used in our experiments and then present the results. For all experiments, we have chosen $m = 2$, $C_{min} = 2$ and $C_{max} = 10$.

5.1 Data sets

In order to provide a variety of data types, we generated four different groups of artificial data sets with controlled OLR between clusters, based on the theory discussed in Section 4. Each data set is represented by X_{g_s}, with g varying from 1 to 4 and s from 1 to 10, i.e., there are 10 data sets in each group. The first group is a collection of 2-dimensional data sets, the second is a collection of 3-dimensional data sets, while the third and fourth groups are in 4- and 5-dimensional space, respectively.

All data sets in a group are generated from a common mixture with slightly different values of the mixture parameters. Specifically, the parameter values for a set X_{g_s} are obtained by modifying the parameter values for the set X_{g_1}. The modification is carried out in such a way that the maximum OLR between any two clusters in a set varies from 0.06 to 0.9 for each group $(X_{g\_1}, \ldots, X_{g\_10})$. We have chosen OLR values within [0.06, 0.9] in order to perform a fair evaluation of the validity indices. The case in which $OLR = 0.06$ corresponds to data sets with only well-separated clusters, whereas the case in which $OLR = 0.9$ corresponds to data sets with strongly overlapped clusters.

Each set in group X_1 has four clusters with a total of 2000 points. Clusters 1 and 3 have a spherical shape, while the two others have an elliptical shape. Each $X_{1\_s}, (s > 1)$ is obtained by successively changing both axes of the second cluster and the orientation of the fourth cluster. Figures 4, 5 and 6 show the data sets X_{1_1}, X_{1_6} and X_{1_10}, for which the maximum OLR between cluster 2, cluster 3 and cluster 4 is 0.06, 0.5 and 0.9 respectively. The main characteristic of the data sets in this group is the continuous change in the covariance structure of clusters, i.e., the shape of the second cluster and the orientation of the fourth cluster. The overlap here is between three clusters,

while we keep the first cluster well separated.

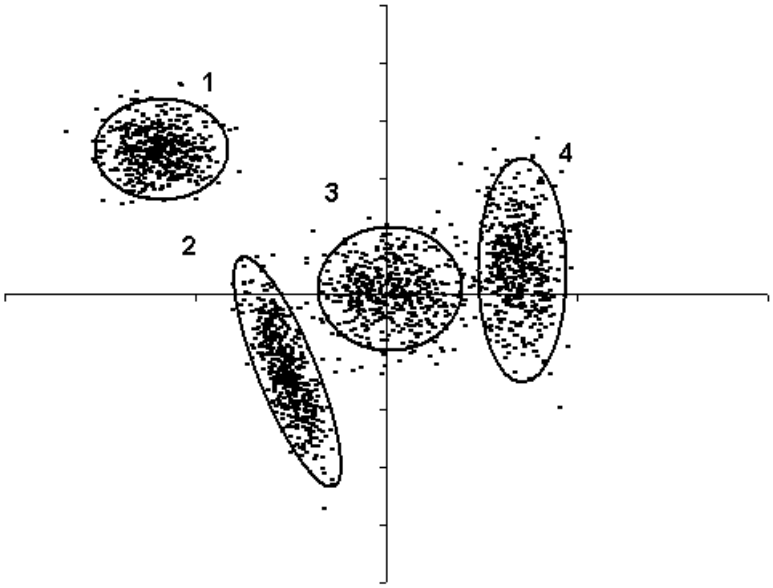

Fig. 4. Data set X_{1_1}; the maximum OLR is 0.06

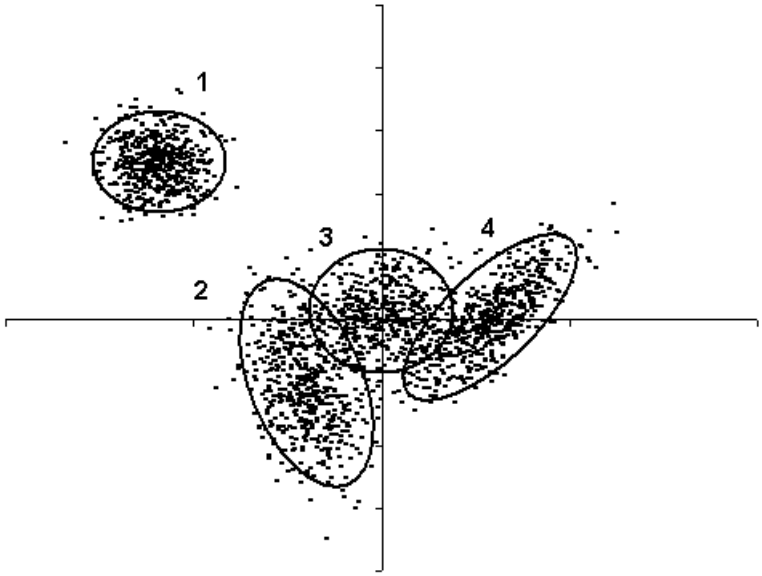

Fig. 5. Data set X_{1_6}; the maximum OLR is 0.5

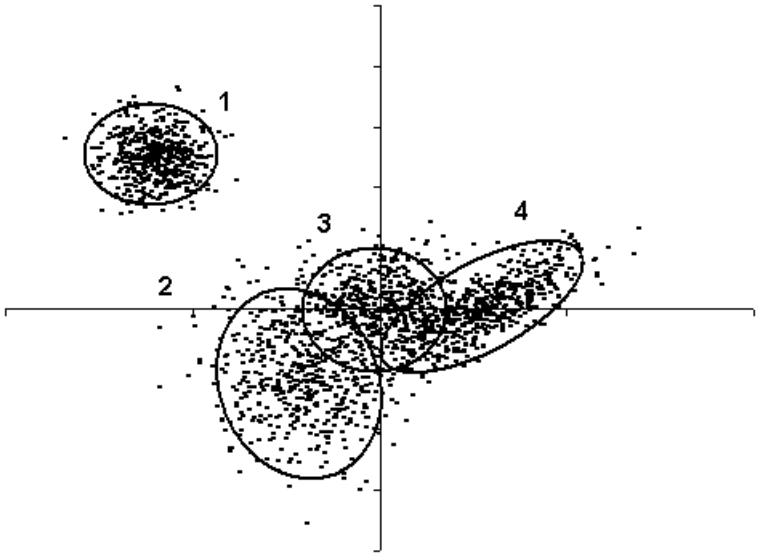

Fig. 6. Data set X_{1_10}; the maximum OLR is 0.9

Each set in group X_2 has five ellipsoidal clusters with a total of 1650 points. The number of data points in each cluster varies. In addition to this, the structure of the clusters differs. Cluster 1 is small and dense, while clusters 2 and 3 are large. Cluster 4 is sparse while cluster 5 is small. Each $X_{2\_s}, (s > 1)$ is obtained by successively approaching cluster 1 to cluster 2 and cluster 4 to cluster 5, and by changing the orientation of cluster 3 with respect to the x-axis. Figures 7, 8 and 9 illustrate X_{2_1}, X_{2_6} and X_{2_10}, respectively. The maximum *OLR* for X_{2_1}, X_{2_6} and X_{2_10} is 0.06, 0.5 and 0.9, respectively. The main characteristic of the data sets in this group is the presence of elliptical clusters with different size and density.

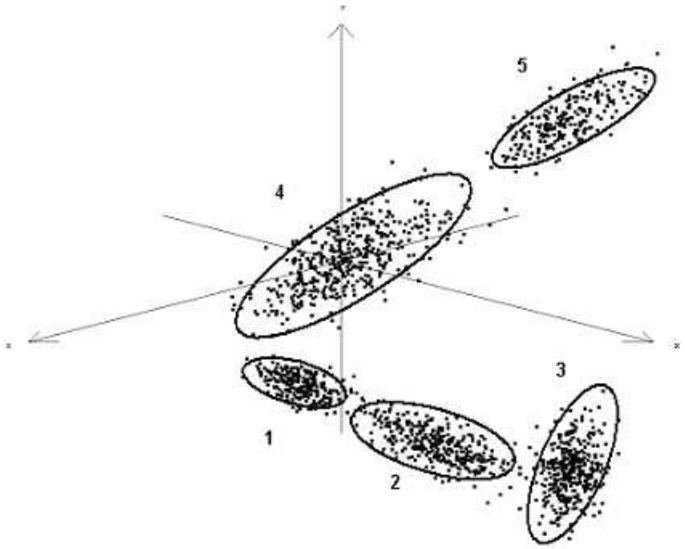

Fig. 7. Data set X_{2_1}; the maximum OLR is 0.06

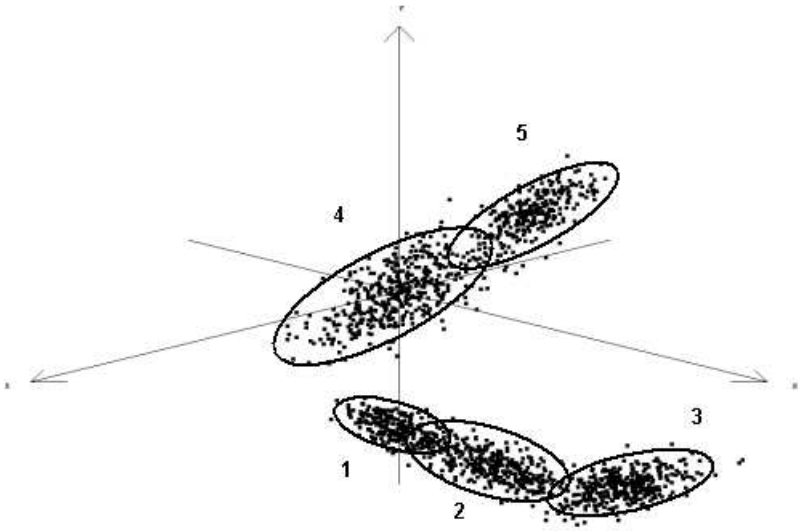

Fig. 8. Data set X_{2_6}; the maximum OLR is 0.5

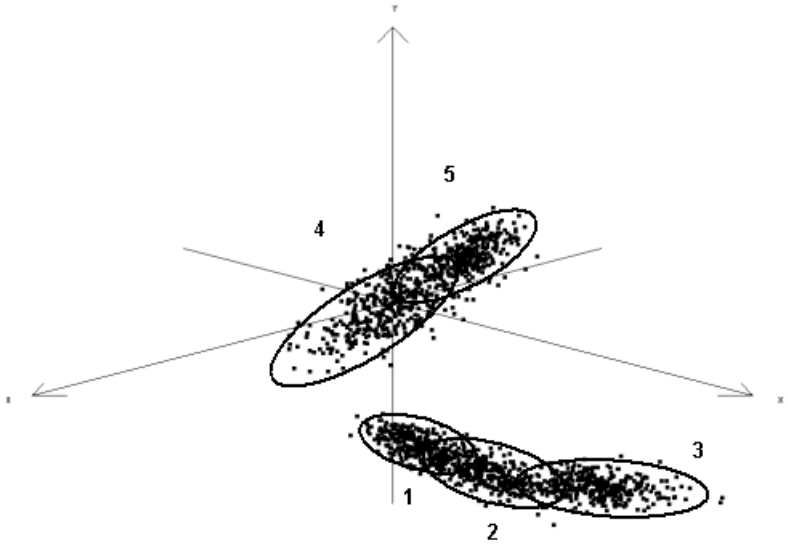

Fig. 9. Data set X_{2_10}; the maximum OLR is 0.9

Each set in the third group X_3 contains six hyper-spherical clusters with a total of 1600 points. There is a significant difference in density between the clusters. Each $X_{3\_s},\ (s > 1)$ is obtained by successively moving clusters 1, 2 and 3 towards each other. We perform the same operation for clusters 4, 5 and 6. The main characteristic of the data sets in this group is the presence of spherical clusters with different density. The overlap here is between three clusters at a time.

In each set of the last group, X_4, there are four clusters with different shapes. Two of them (clusters 1 and 4) are hyper-ellipsoidal and the two others (clusters 2 and 3) are hyper-spherical (the total number of data points in each set is 200). There are some significant differences in geometric shape between the two ellipsoidal clusters. Cluster 1 is sparse, while cluster 4 is long and narrow. Each $X_{4\_s},\ (s > 1)$ is obtained by successively moving cluster 1 towards cluster 2 and cluster 3 towards cluster 4. The main characteristic of the data sets in this group is the presence of large clusters with different shape and low density.

5.2 Experimental results

The tables below summarize the results of all of the validity indices introduced in Section 3 in conjunction with the FMLE algorithm, for the four groups of data sets X_1, X_2, X_3 and X_4.

Table 2
The optimal number of clusters using FMLE for X_1

Data Set	OLR	V_{PC}	V_{FS}	V_{XIE}	V_{ZLE}	V_{N_INV}	V_{XRZ}	V_{WSJ}	V_{FH}	V_{APD}	V_{SC}
X_{1_1}	0.06	2	**4**	3	2	2	3	**4**	**4**	**4**	**4**
X_{1_2}	0.10	2	**4**	3	2	2	3	**4**	**4**	**4**	**4**
X_{1_3}	0.20	2	**4**	2	2	2	2	**4**	**4**	**4**	**4**
X_{1_4}	0.30	2	**4**	2	2	2	2	**4**	**4**	**4**	**4**
X_{1_5}	0.40	2	**4**	2	2	2	2	**4**	**4**	**4**	**4**
X_{1_6}	0.50	2	**4**	2	2	2	2	**4**	**4**	3	**4**
X_{1_7}	0.60	2	**4**	2	2	2	2	**4**	3	3	**4**
X_{1_8}	0.70	2	**4**	2	2	2	2	**4**	3	2	**4**
X_{1_9}	0.80	2	**4**	2	2	2	2	**4**	3	2	**4**
X_{1_10}	0.90	2	**4**	2	2	2	2	3	2	2	**4**

Table 3
The optimal number of clusters using FMLE for X_2

Data Set	OLR	V_{PC}	V_{FS}	V_{XIE}	V_{ZLE}	V_{N_INV}	V_{XRZ}	V_{WSJ}	V_{FH}	V_{APD}	V_{SC}
X_{2_1}	0.06	2	7	4	2	3	**5**	**5**	**5**	**5**	**5**
X_{2_2}	0.10	2	7	4	2	3	**5**	**5**	**5**	**5**	**5**
X_{2_3}	0.20	2	6	4	2	2	**5**	**5**	**5**	**5**	**5**
X_{2_4}	0.30	2	7	4	2	2	**5**	**5**	**5**	**5**	**5**
X_{2_5}	0.40	2	7	4	2	2	**5**	**5**	**5**	4	**5**
X_{2_6}	0.50	2	**5**	2	4	2	4	**5**	3	4	**5**
X_{2_7}	0.60	2	7	4	2	2	4	4	3	4	**5**
X_{2_8}	0.70	2	7	2	2	2	4	4	3	4	**5**
X_{2_9}	0.80	2	7	2	2	2	3	7	3	4	**5**
X_{2_10}	0.90	2	**5**	2	2	2	3	4	3	4	**5**

Table 4
The optimal number of clusters using FMLE for X_3

Data Set	OLR	V_{PC}	V_{FS}	V_{XIE}	V_{ZLE}	V_{N_INV}	V_{XRZ}	V_{WSJ}	V_{FH}	V_{APD}	V_{SC}
X_{3_1}	0.06	2	**6**	2	**6**	**6**	**6**	**6**	**6**	**6**	**6**
X_{3_2}	0.10	2	**6**	2	**6**	**6**	**6**	**6**	**6**	**6**	**6**
X_{3_3}	0.20	2	**6**	2	**6**	**6**	**6**	**6**	**6**	**6**	**6**
X_{3_4}	0.30	2	**6**	2	5	**6**	4	**6**	**6**	**6**	**6**
X_{3_5}	0.40	2	**6**	2	**6**	**6**	2	4	**6**	**6**	**6**
X_{3_6}	0.50	2	**6**	2	**6**	**6**	2	2	**6**	**6**	**6**
X_{3_7}	0.60	2	**6**	2	**6**	**6**	2	2	**6**	**6**	**6**
X_{3_8}	0.70	2	**6**	2	2	**6**	2	2	2	**6**	**6**
X_{3_9}	0.80	2	**6**	2	2	**6**	2	2	2	**6**	**6**
X_{3_10}	0.90	2	**6**	2	2	5	2	2	2	**6**	**6**

Table 5
The optimal number of clusters using FMLE for X_4

Data Set	OLR	V_{PC}	V_{FS}	V_{XIE}	V_{ZLE}	V_{N_INV}	V_{XRZ}	V_{WSJ}	V_{FH}	V_{APD}	V_{SC}
X_{4_1}	0.06	**4**	**4**	**4**	**4**	2	**4**	**4**	10	9	**4**
X_{4_2}	0.10	2	**4**	**4**	**4**	2	**4**	**4**	10	9	**4**
X_{4_3}	0.20	2	**4**	2	2	2	**4**	**4**	10	5	**4**
X_{4_4}	0.30	2	**4**	2	2	2	2	**4**	10	8	**4**
X_{4_5}	0.40	2	**4**	2	2	2	2	3	10	**4**	**4**
X_{4_6}	0.50	2	8	2	2	2	2	**4**	10	5	**4**
X_{4_7}	0.60	2	9	2	2	2	2	**4**	10	10	**4**
X_{4_8}	0.70	2	8	2	2	2	2	2	10	10	**4**
X_{4_9}	0.80	2	8	2	2	2	2	2	10	8	3
X_{4_10}	0.90	2	10	2	2	2	2	3	10	5	2

The values in the tables above invite several comments.

(1) In all of the experiments, our new validity index, V_{SC}, demonstrates the stability to provide the correct number of clusters for different values of *OLR*.

(2) The behavior of the partition coefficient is almost the same in all cases. It tends to fail even when OLR is low. In general, we can say that V_{PC} is very sensitive to overlapping clusters. It performs well only with well-separated spherical clusters.
(3) The behavior of V_{FS} is special in that it reacts differently from all the other indices tested to increasing OLR. It tends to favor a larger number of smaller clusters (Tables 3 and 5). Although it performs very well with X_1 and X_3 (Tables 2 and 4), it fails completely with X_2 (Table 3). Overall, the experiments reported here seem to suggest that, in general, V_{FS} provides good results with data sets that include overlapping clusters.
(4) From the experiments on the four generated groups of data sets, we can see that V_{XIE} yields acceptable results when OLR is low. When OLR increases, it tends to favor smaller numbers of clusters. Moreover, the experiments show that this index fails even for data sets with very small OLR values (0.06 and 0.1). It appears that elongated cluster shape and sparse density have a negative impact on the results yielded by V_{XIE}.
(5) V_{ZLE} and V_{N_INV} perform very well with X_3, whereas they fail completely with X_1, X_2 and X_4. From the results of the experiments on the four groups, we can say that V_{ZLE} and V_{N_INV} perform well in situations involving overlapping clusters with spherical shape. However, it would appear that wide variations in cluster covariance structure can greatly influence the results yielded by these indices.
(6) In general, with small values of OLR, V_{XRZ} yields good results. However, as OLR increases, V_{XRZ} gradually loses its ability to distinguish between overlapping clusters. In addition to this, we remark that this index fails with X_1. We conclude that wide variation in cluster structure has a negative influence on the results of V_{XRZ}.
(7) From the experiments on the first three groups of data sets, we remark that V_{FH} and V_{APD} are able to find the true number of clusters when $OLR \leq 0.6$. However, with higher values of OLR these indices encounter difficulties. In addition to this, V_{FH} and V_{APD} fail completely with the last group, X_4. We believe that the presence of large clusters with low density adversely affects the results of V_{FH} and V_{APD}.
(8) From the experiments on the four generated groups of data sets, we can see that V_{WSJ} provides good results when $OLR \leq 0.6$. However, higher values of OLR can influence the performance of this index.

5.3 Experiments on real-world data sets

The suitability of the validity indices was also tested on three real-world data sets containing separate and overlapping clusters. The first data set is Haberman's survival data (Haberman), which contains 306 data points with 3 features, from two well-separated clusters. The second is the Iris data set

(Iris), widely used for testing validity indices. It consists of three clusters, each of which contains 50 observations. Of these three clusters, two are overlapped. These first two data sets are from the UCI machine learning repository (http://www.ics.uci.edu/mlearn/MLSummary.html). The last data set is the Crude Oil data set (Oil) (Johson and Wichern, 1998). Crude oil samples were analyzed from three zones of sandstone: Wilhelm, Sub-Mulinia, and Upper. This data set contains 56 data points with 5 features, from three overlapping clusters. The numbers of clusters yielded by all the validity indices for the three real-world data sets are given in Table 6.

Table 6
The optimal number of clusters for three real-world data sets

Data Set	V_{PC}	V_{FS}	V_{XIE}	V_{ZLE}	V_{N_INV}	V_{XRZ}	V_{WSJ}	V_{FH}	V_{APD}	V_{SC}
Haberman	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Iris	2	**3**	2	2	2	2	**3**	**3**	2	**3**
Oil	2	4	2	2	2	2	**3**	10	10	**3**

As can be seen from the table, all the validity indices perform well with the Haberman data. The good reliability of all the validity indices is due to the fact that the Haberman data set is composed of two well-separated spherical clusters. However, if a data set contains overlapping clusters, some validity indices encounter difficulties. In the case of the Iris data, V_{FS}, V_{WSJ}, V_{FH} and V_{SC} yield the correct number of clusters, while the other validity indices fail to do so. In the case of Oil, which is a small data set containing three overlapping clusters, only V_{WSJ} and V_{SC} are able to find the true number of clusters. V_{FH} and V_{APD} favor a large number of small clusters. We believe that the low density in each of the three clusters influences the results yielded by these indices. Due to the overlap between the three clusters in the Oil data set, V_{PC}, V_{XIE}, V_{ZLE}, V_{N_INV} and V_{XRZ} yield two clusters, while V_{FS} yields four clusters. Similar behavior was observed for V_{FS} in the generated data sets (see comment 3 in Section 5.2)

6 Conclusion

We have proposed a new validity index for fuzzy clustering and a novel method for performing comparison and evaluation of validity indices in relation to cluster overlap. The effectiveness of our new index in coping with cluster overlap and shape and density variation was demonstrated experimentally for a number of generated data sets. The reliability of the proposed index was also verified on three real-world data sets.

In order to compare the capacity of cluster validity indices to distinguish

between overlapping clusters, an extensive evaluation was carried out with truthed data sets generated on the basis of the theory of overlap developed by Sun and Wang (2003). The study reveals that although all the indices work well with well-separated spherical clusters, only a few perform well with overlapped clusters. These indices should receive more consideration if separating partially overlapped clusters is one of the essential features of a clustering system. The concept of overlap rate provides a unified and objective measure of the difficulty of separating clusters in a data set. It serves as a good indicator of the performance of clustering algorithms and validity indices.

Acknowledgments
We gratefully thank anonymous reviewers for their many helpful and constructive comments and suggestions. This work has been supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Aitnouri, E. et al., 2000. On Clustering Techniques Comparison For Histogram pdf Estimation. Pattern Recognition and Image Analysis 10 (2), 206–217.

Bezdek, J.C., Dunn, J.C., 1975. A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions. IEEE Transactions on Computers 24 (8), 835–838.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York.

Bezdek, J.C. et al., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, Boston.

Duda, R. O. et al., 2001. Pattern Classification (2nd ed.). John Wiley & Sons, New York.

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well-Separated Clusters. Journal of Cybernetics 3 (3), 32–57.

Fukunaga, K., 1990. Introduction to Statistical Pattern Regognition (2nd ed.). Academic Press Professional, San Diego, CA.

Gath, I., Geva, A.B., 1989. Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7), 773–781.

Geva, A.B. et al., 2000. A Comparison of Cluster Validity Criteria for a Mixture of Normal Distributed Data. Pattern Recognition Letters 21 (6-7), 511-529.

Hoppner, F. et al., 1999. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons, England.

Johson, R.A., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis (4th ed.). Prentice Hall, New York.

Krishnapuram, R., Keller, J., 1993. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems 1 (2), 98–110.

Maulik, U., Bandyopadhyay, S., 2002. Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12), 1650–1654.

Pal, N. R., Bezdek, J.C., 1995. On Cluster Validity for the Fuzzy C-Means Model. IEEE Transaction on Fuzzy Systems 3 (3), 370–379.

Pal, N. R., Biswas, J., 1997. Cluster Validation using Graph Theoretic Concepts. Pattern Recognition 30 (6), 847–857.

Rezae, M. et al., 1998. A New Cluster Validity Index for the Fuzzy C-Mean. Pattern Recognition Letters 19 (3-4), 237–246.

Sun, H., Wang, S., 2003. Overlap Degree Between Two Components in Mixture Models. Research Report no. 345, University of Sherbrooke.

Sun, H. et al., 2004. FCM-based Model Selection Algorithms for Determining the Number of Clusters. Pattern Recognition 37 (10), 2027–2037.

Titterington, D. et al., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons, New York.

Xie, X.L., Beni, G., 1991. A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (8), 841–847.

Xie, Y. et al., 2002. 3M Algorithm: Finding an Optimal Fuzzy Cluster Scheme for Proximity Data. In: Proceedings of the FUZZ-IEEE Conference- 2002 IEEE World Congress on Computational Intelligence, Honolulu, HI.

Zahid, N. et al., 1999. A New Cluster Validity for Fuzzy Clustering. Pattern Recognition 32 (7) 1089–1097.