Data Min Knowl Disc (2011) 23:479–502
DOI 10.1007/s10618-011-0212-3

# Measuring the component overlapping in the Gaussian mixture model

**Haojun Sun · Shengrui Wang**

Received: 20 May 2009 / Accepted: 27 January 2011 / Published online: 27 February 2011


**Abstract** The ability of a clustering algorithm to deal with overlapping clusters is a major indicator of its efficiency. However, the phenomenon of cluster overlapping is still not mathematically well characterized, especially in multivariate cases. In this paper, we are interested in the overlap phenomenon between Gaussian clusters, since the Gaussian mixture is a fundamental data distribution model suitable for many clustering algorithms. We introduce the novel concept of the ridge curve and establish a theory on the degree of overlap between two components. Based on this theory, we develop an algorithm for calculating the overlap rate. As an example, we use this algorithm to calculate the overlap rates between the classes in the IRIS data set and clear up some of the confusion as to the true number of classes in the data set. We investigate factors that affect the value of the overlap rate, and show how the theory can be used to generate "truthed data" as well as to measure the overlap rate between a given pair of clusters or components in a mixture. Finally, we show an example of application of the theory to evaluate the well known clustering algorithms.

**Keywords** Mixture model · Ridge curve · Overlap rate · Cluster analysis

Responsible editor: Charu Aggarwal.

The expression "simulation data" is used in this paper to designate a data set with known membership of data points w.r.t. each cluster.

H. Sun (✉)
College of Engineering, Shantou University, Shantou 515063, Guangdong, China
e-mail: haojunsun@stu.edu.cn

S. Wang
Département d'informatique, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada

## 1 Introduction

The Gaussian mixture plays an important role in cluster analysis. It is a basic data model to EM-based clustering algorithms as well as to some advanced partition-based algorithms such as the hybrid Fuzzy Maximum Likelihood Estimation algorithm (FMLE) (Gath and Geva 1989). On the other hand, the simplified Gaussian mixture in which each component has a spherical shape is also frequently used as the data model for popular algorithms such as K-means and the Fuzzy C-means (Bezdek 1981), though it is not the only data model suitable to these algorithms. All these clustering algorithms or their improved versions are widely used in practice (Ramos and Muge 2000; Hsu 2000; Nicholls and Tudorancea 2001; Fraley 1998). However, their performance often depends on whether the data set contains well separated clusters, or in other words, whether and how the components of the Gaussian mixture overlap each other.

Given a data set satisfying the distribution of a mixture of Gaussians, the degree of overlap between components affects the number of clusters "perceived" by a human operator or detected by a clustering algorithm. In other words, there may be a significant difference between intuitively detected clusters and the true clusters corresponding to the components in the mixture. The component overlapping phenomenon is illustrated in Fig. 1. Figure 1a–c show the 1-D case, with two components that are (almost) non-overlapping, partially overlapping and strongly overlapping. Fig. 1d–f show their counterparts in the 2-D case. Non-overlapping clusters are relatively easy to be discovered by clustering algorithms. Partially overlapped clusters are more difficult to separate and strongly overlapping clusters are in general very difficult to separate whichever clustering algorithm is used. Strictly speaking, if we assume that a cluster corresponds exactly to a component in the mixture, it is then necessary for the clustering algorithm to separate overlapped clusters. In practice, however, whether it is desirable to separate overlapped clusters depends on the application. Color image

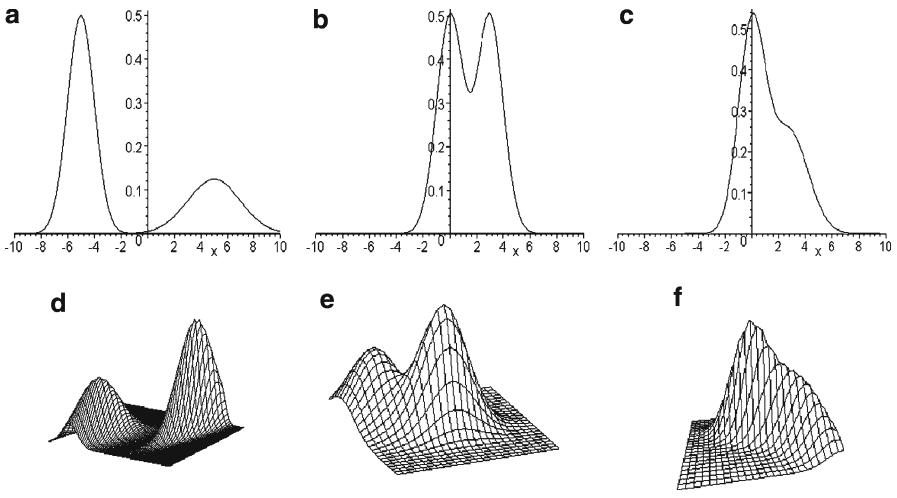


**Fig. 1** Two components that are non-, partially and strongly overlapping, in the 1-D and 2-D cases

segmentation is a good example in which (partially) overlapped clusters in the pixels' color space may represent either a single object or several related objects.

Cluster overlap impacts on another important issue of clustering, which is determining the number of clusters. Existing methods, often based on measuring the validity of each possible number of clusters, yield good estimates when the clusters are well separated (non-overlapping). However, for data sets with overlapping clusters, the results are often unpredictable. Examples of such results are reported in our recent paper on determining the number of clusters using the Fuzzy C-means algorithm (Sun et al. 2004). One of the main reasons for this problem is that many algorithms fail to distinguish between partially overlapping clusters.

We are interested in establishing analytically the relationship between the degree of overlap and the parameters of each component. This relationship will have an important impact in many real applications. For example, simulation data sets with a prescribed degree of overlap between clusters provides a way of evaluating the capacity of existing clustering algorithms to identify overlapping clusters. In image processing, the degree of overlap between two objects (clusters) in a color image measures the similarity of the objects. The magnitude of the overlap rate provides a useful indicator on whether the two objects should be considered the same. The overlap rate can also be used as a similarity measure for deciding merging of clusters.

The main contributions of this paper are creating a theoretical framework for component overlap in a mixture, proposing a practical algorithm for measuring the degree of overlap between two components, and describing its application to evaluate the well known clustering algorithms. The concepts and the algorithm developed here can be readily used to study the overlap phenomenon in a multi-component mixture by measuring the degree of overlap between any pair of clusters.

We do not make any assumption regarding the covariance structure of each component. For the sake of simplicity, the theory will be introduced in the 2-D case. All of the results hold in the multidimensional case and the main equations for the multidimensional case are discussed. The theory is based on a novel concept, that of the "ridge curve". A series of theorems will be established to show the main characteristics of component overlapping. This allows us to give a feasible definition of the overlap rate and to develop algorithms for measuring it. The theory also allows us to effectively generate simulation data sets with user-defined overlap rate.

This paper is organized as follows. In the next section, we give a theoretical framework for measuring the similarity between two components in a Gaussian mixture, introduce the concept of the "ridge curve", and prove a series of theorems to establish the theoretical framework for describing the phenomenon of overlapping components. In Sect. 3, we define the overlap rate between two components in a mixture and design an algorithm for calculating this overlap rate. As an example, we examine overlapping in the IRIS data set (the data set consists of 150 data points with four features and three classes; there are 50 data points per class). We also show why the more intuitive interpretation of overlap based on the classification error is not a good choice. In Sect. 4, we consider the factors affecting the value of the overlap rate in order to generate simulation data sets with prescribed overlap rates. And we make use of the component overlap theory to develop a new method to evaluate clustering algorithms. We conclude our paper with a discussion of some extensions of the research.

This paper is extended from the conference paper (Sun and Wang 2004) where the idea of the overlapping rate was initially introduced. In the current manuscript, in addition to providing more articulated motivation, more detailed explanation of the concept as well as analysis of related work, we have compared the propose method for computing overlap rate to several existing methods for measuring the "overlapping". Especially, we showed why the overlap rate is a better measure than the classification error rate. Furthermore, we have extended the overlap rate concept to multiple-dimensional data and developed an algorithm for the general case, and finally applied the new theory to compare clustering algorithms based on the generated data sets with overlapping phenomenon.

## 2 Theoretical framework

### 2.1 Mixture models

Mixture models satisfy some intuitive definitions of cluster structure. Two key properties of a cluster are internal cohesion, which requires that entities within the same cluster should be similar to each other, and external isolation, which requires that entities in one cluster should be separated from entities in another cluster by fairly empty areas of space (Milligan 1980). Internal cohesion is an inherent property of mixture models, since for a given cluster, data are generated from the same distribution. External isolation concerns the degree of overlap between the components of the mixture model (Aitnouri et al. 2002).

A set of $n$ entities forming a $d$-dimensional data can be presented as $\mathbf{X} = \{X_1, ..., X_n\}$, where $X_i$ is a vector of dimension $d$. In the finite mixture models dealt with here, each $X_i$ can be viewed as arising from a mixture of $k$ Gaussian distributions and the probabilistic density function (*pdf*) is given (in the $d$-dimensional data space) by:

$$\begin{cases} p(X) = \sum_{i=1}^{k} \alpha_i G_i(X, \mu_i, \Sigma_i) \\ X \in R^d \end{cases} \tag{1}$$

with the restrictions $\alpha_i > 0$ for $i = 1, ..., k$ and $\sum_{i=1}^{k} \alpha_i = 1$. $(\mu_i, \Sigma_i)$ denote, respectively, the mean and the covariance matrix for the $i$th distribution $G_i$. $G_i$ is the $i$th component, given by:

$$G_i(X) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1}(X - \mu_i)\right) \tag{2}$$

In this paper, our investigation of the overlap phenomenon follows a geometrical approach which is intuitive and practical in real applications. Because of complex non-linear equations involved and of the real need in practice, the current theoretical study is restrained to the overlap of two components. The overlap phenomenon of three or more components could be dealt with using our theory in a pair-wise way.

Without loss of generality, we assume, in the development of our theory, that $k = 2$ in Eq. 1. If the two components are (almost) non-overlapping or only partially overlapping, then the *pdf* of the mixture has two separate local peaks. In this case, the data arising from the mixture model are viewed as two clusters (see Fig. 1a, b, d, e). On the other hand, if the *pdf* has only one peak, then the data are viewed as having two strongly overlapping components. Visually, it is very difficult to distinguish between the two components in this case (see Fig. 1c, f).

### 2.2 Related work

Most of the methods used to measure the degree of overlap between two components are probabilistic in nature. A few others are geometric. Probabilistic methods are based on a hypothesis that the data are drawn from one of several probabilistic distributions. The mathematical model is a mixture of these probabilistic distributions. The most commonly used model is the Gaussian mixture. The classification error rate, $\int \min\{\alpha_1 G_1, \alpha_2 G_2\} dX$, as well as various distances such as the Mahalanobis distance (Day 1969; McLachlan and Basford 1988), $D_{Mah} = ((\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2))^{1/2}$, and an extension of it, the Bhattacharyya distance (Fukunaga 1990; Do and Vetterliyx 2000), $D_{batt} = \frac{1}{8}(\mu_1 - \mu_2)^T [\frac{\Sigma_1 + \Sigma_2}{2}]^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}}$ are used as measures of the similarity between two Gaussian distributions (Chan et al. 2003; Salvi 2003). Because of the difficulty of computing the classification error rate, it is often replaced by its upper bounds [for example, the Bhattacharyya bound (Fukunaga 1990), $B_{batt} = \sqrt{\alpha_1 \alpha_2} e^{-D_{batt}}$] in practical applications. Another type of distance is defined directly by their *pdf*: an example is the Kullback–Leibler distance (Kullback 1959), $D_{kl} = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx$. However, this type of distance is not symmetric, and is difficult to use in real applications.

Distance measures fail to characterize component overlap for various reasons. For instance, the distance between the centers of components, $|\mu_1 - \mu_2|$, ignores the covariance matrix and other probability parameters. The Mahalanobis distance measures the similarity between two normal distributions, assuming that they have the same covariance matrix. The Bhattacharyya distance takes covariance matrices into account, but ignores the influence of the coefficients of the components. We will provide cases in the Sect. 3.5 in which the classification error rate shows behaviors that are clearly not consistent with the changing phenomenon of overlap. Its main problem is that it does take account explicitly of the geometrical properties of the *pdf*.

Geometric methods consider the geometrical properties of the *pdf* of a mixture. The typical approach is based on comparing between the local minimum and the local maximum of *pdf* or their locations. In Tabbone (1994), Tabbone studied the presence of a false edge between the steps of a staircase edge. He treated this as a mixture of two Gaussians with equal covariance, $\sigma$, in the 1-D case. There is a false edge (the two Gaussians strongly overlap) if and only if $2\sigma < |\mu_1 - \mu_2|$. In Aitnouri et al. (2002), Aitnouri et al. extended the results in Tabbone (1994) to a mixture of Gaussians with non-equal covariance. The extension remained restricted to the 1-D case. In Aitnouri et al. (2002), the overlap rate was defined as the ratio of the height of the minimum

between the two component centers (if such a minimum exists) to the height of the lower maximum also situated between the component centers. This definition does not have a natural extension to the multi-dimensional case since there is no local minimum between the two component centers.

In what follows, we try to characterize this phenomenon as a function of the mixture parameters. In particular, we derive an efficient procedure for verifying whether the two components strongly overlap and to compute an overlap rate when they partially overlap. For the sake of simplicity, we establish the theory for the 2-D case ($d=2$). All of the theorems are also valid in the multi-dimensional case with an appropriate definition of the ridge curve, the key concept of this theory, which will be discussed in the following subsection. There, we will prove that the peaks of the *pdf* in $R^2$ can be found by a search procedure which follows a curve linking the centers of the two components.

### 2.3 Peaks of the *pdf* and ridge curve

If the *pdf* of a mixture has two separated local peaks, then the overlap rate between the two components depends on the geometrical properties of the two peaks. In this case, there is a saddle between the two peaks, and the value of the saddle is less than the value of either peak. We propose to use the ratio of the saddle to the lower peak to measure the degree of overlap between the two components. This idea relates to the results described in Aitnouri et al. (2002). The challenge in the multi-dimensional case is how to compute the overlap rate.

As we know, the peaks of the *pdf* satisfy the following system of stationary equations:

$$\begin{cases} \frac{\partial p}{\partial x_1} = A_{x_1}\alpha_1 G_1 + B_{x_1}\alpha_2 G_2 = 0 & (I) \\ \frac{\partial p}{\partial x_2} = A_{x_2}\alpha_1 G_1 + B_{x_2}\alpha_2 G_2 = 0 & (II) \end{cases} \tag{3}$$

where

$$\begin{aligned} \begin{pmatrix} A_{x_1} \\ A_{x_2} \end{pmatrix} &= \nabla \|X-\mu_1\|^2_{\Sigma_1^{-1}} = -\Sigma_1^{-1}(X-\mu_1) \\ \begin{pmatrix} B_{x_1} \\ B_{x_2} \end{pmatrix} &= \nabla \|X-\mu_2\|^2_{\Sigma_2^{-1}} = -\Sigma_2^{-1}(X-\mu_2) \end{aligned} \tag{4}$$

Because of the involvement of $G_1$ and $G_2$ in Eq. 3, this system does not have a closed-form solution. A naive numerical solution would imply searching a whole region of $R^2$ ($R^d$ in the general case). The following theorems illustrate that the search procedure can be restricted to a curve.

**Theorem 1** *A general mixture model of two Gaussian distributions, given by Eq. 1, can be converted to two special forms by implementing an affine transformation on $X$. The two special forms are given by:*

$$\begin{cases} \mu_1 = (\mu_{11}, \mu_{12})^T = (0,0)^T, \quad \mu_2 = (\mu_{21}, \mu_{22})^T \\ \Sigma_1 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} \rho_1^2 & 0 \\ 0 & \rho_2^2 \end{pmatrix} \end{cases} \tag{5}$$

*and*

$$\begin{cases} \mu_1 = (\mu_{11}, \mu_{12})^T = (0,0)^T, \quad \mu_2 = (\mu_{21}, 0)^T \\ \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} \rho_{11}^2 & \rho_{12} \\ \rho_{12} & \rho_{22}^2 \end{pmatrix} \end{cases} \tag{6}$$

In other words, the two covariance matrices are diagonal and one component is centered at the origin. The proof of this theorem is based on simultaneous diagonalization; see Fukunaga (1990).

Now we introduce a new concept, that of the ridge curve (RC) of a mixture model of two Gaussian distributions, as follows:

**Definition 1** Given a mixture of two Gaussians (Eq. 1), the quadratic curve

$$A_{x_1} B_{x_2} - B_{x_1} A_{x_2} = 0 \tag{7}$$

is called the RC of the mixture. Here, $A_{x_1}$, $B_{x_2}$, $B_{x_1}$ and $A_{x_2}$ are defined by Eq. 4.

**Theorem 2** *The means of the two components and the stationary points (peak points and saddle points) of the pdf, $p(X, \mu_1, \mu_2, \Sigma_1, \Sigma_2, \alpha_1, \alpha_2)$, are on the ridge curve.*

*Proof* The first part of this theorem can be easily proved by considering that both $A_{x_1} B_{x_2}$ and $B_{x_1} A_{x_2}$ contain the factors $(X - \mu_1)$ and $(X - \mu_2)$.

For the second part, we mentioned above that any stationary point $(x_1, x_2)^T$ of the *pdf* should satisfy Eq. 3. The stationary equation can be written as:

$$\begin{pmatrix} A_{x_1} & B_{x_1} \\ A_{x_2} & B_{x_2} \end{pmatrix} \begin{pmatrix} \alpha_1 G_1(X, \mu_1, \Sigma_1) \\ \alpha_2 G_2(X, \mu_2, \Sigma_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{8}$$

We know that for any $X = (x_1, x_2)^T$, the vector $\begin{pmatrix} \alpha_1 G_1(X, \mu_1, \Sigma_1) \\ \alpha_2 G_2(X, \mu_2, \Sigma_2) \end{pmatrix}$ is non-zero. Thus the matrix $\begin{pmatrix} A_{x_1} & B_{x_1} \\ A_{x_2} & B_{x_2} \end{pmatrix}$ is singular. This means $\begin{vmatrix} A_{x_1} & B_{x_1} \\ A_{x_2} & B_{x_2} \end{vmatrix} = 0$ or $A_{x_1} B_{x_2} - B_{x_1} A_{x_2} = 0$. Thus the stationary points of the *pdf* are on the *RC*. ◇

Based on Theorem 2, we conclude that it is sufficient to search the ridge curve to find the stationary points of the *pdf*. However, due to the quadratic nature of the curve, the search procedure must solve the alternative problem, in addition to restraining the search interval. Indeed, the following theorem will indicate that all of the stationary points of the *pdf* and the means of the components fall on the same segment of the ridge curve.

**Theorem 3** *The stationary points of the pdf described by Eq. 1 fall on the segment between the two means of the components of the ridge curve.*

For the proof of this theorem, we need to introduce some lemmas.

**Lemma 1** *Based on the special case mentioned in Theorem 1, the ridge curve,* $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$*, is a hyperbola or a line.*

*Proof* Let $b = \frac{1}{\sigma_1^2\rho_2^2} > 0$ and $c = -\frac{1}{\sigma_2^2\rho_1^2} < 0$. The curve $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$ can be written as $(b+c)x_1x_2 - b\mu_{22}x_1 - c\mu_{21}x_2 = 0$.

If $b + c = 0$ or if $\mu_{22}\mu_{21} = 0$, then the above equation degenerates to a line. If $b + c \neq 0$ and $\mu_{22}\mu_{21} \neq 0$, implementing the following affine transformation to $X = (x_1, x_2)^T$,

$$\begin{cases} x_1^{'} = x_1 - \frac{c\mu_{21}}{b+c} \\ x_2^{'} = (b+c)x_2 - b\mu_{22} \end{cases} \tag{9}$$

we get the normalized hyperbolic curve: $x_1^{'}x_2^{'} = \frac{bc\mu_{12}\mu_{22}}{b+c}$. ◇

**Lemma 2** *Under the same conditions as for Lemma 1, the stationary points of Eq. 1 are inside the rectangle defined by the following four points:* $((0,0)^T, (0,\mu_{22})^T, (\mu_{21}, \mu_{22})^T, (\mu_{21}, 0))^T$.

*Proof* Suppose that a stationary point $X_0 = (x_1^0, x_2^0)^T$ is outside the rectangle. Without loss of generality, we suppose that $\mu_{21} > 0$ and $X_0$ is to the right of the rectangle, i.e. $x_1^0 > \mu_{21}$. Consider the radial

$$\begin{cases} x_1 = \mu_{21} + t(t > 0) \\ x_2 = x_2^0 \end{cases} \tag{10}$$

Both $(X-\mu_1)\Sigma_1^{-1}(X-\mu_1)^T) = X\Sigma_1^{-1}X^T$ and $(X-\mu_2)\Sigma_2^{-1}(X-\mu_2)^T)$ are monotonically increasing. Thus the value of $p(X)$ is monotonically decreasing on the radial. This means that $X_0$ is not a stationary point of the *pdf*.

Here, we note the following fact: let $xy = a$ denote a hyperbola in $R^2$, and $(x_1, y_1)^T$ and $(x_2, y_2)^T$ be two points on the hyperbola. The two points are on the same branch iff $x_1x_2 > 0$ or $y_1y_2 > 0$.

Combining Theorem 2 and Lemmas 1 and 2, we obtain the proof of Theorem 3 as follows.

*Proof of Theorem 3* According to Theorem 1, affine transform $X = mZ + N$ ($m$ is a $2 \times 2$ affine matrix and $N$ is 2-D vector), $Z = (z_1, z_2)^T \in R^2$, transforms the mixture to the special case mentioned in Theorem 1. The ridge curve $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$ is transformed to $A_{z_1}B_{z_2} - B_{z_1}A_{z_2} = 0$ and the stationary points of the *pdf* about $X$ are translated to the stationary points about $Z$, because the transform is affine. Based on Lemmas 1 and 2, the stationary points of the *pdf* about $Z$ fall on the segment between the two means of the components of the curve, $A_{z_1}B_{z_2} - B_{z_1}A_{z_2} = 0$. Thus the stationary points of the *pdf* described by Eq. 1 fall on the segment between the two means of the components of the curve, $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$. ◇

Based on the result of Theorem 3, to determine the stationary points of Eq. 1, we need only search the segment of the $RC$, $A_{x_1} B_{x_2} - B_{x_1} A_{x_2} = 0$, between the two means. The search for the stationary points of a Gaussian mixture is thus reduced to a linear search procedure.

## 3 Degree of overlap

In this section, we will give the definition of the *overlaprate* (*OLR*) between two components of a mixture and design an algorithm to compute the *OLR*. To provide some further insight into the *OLR*, we apply the algorithm to the well-known IRIS data set. Finally, we compare the *OLR* with the classification error rate.

### 3.1 Definition of the overlap rate

We propose a geometric method for determining the overlap rate between two Gaussians. In general, a definition for the overlap rate between two Gaussian components implements the following principle: (1) the value of the overlap rate lies within a normalized interval such as [0, 1]; (2) the overlap rate tends to decrease ($\rightarrow 0$) as the two components become more widely separated; (3) the overlap rate increases ($\rightarrow 1$) as the two components become more strongly overlapped. As mentioned previously, the following definition of overlap rate relies on the ridge curve concept developed in the previous section.

**Definition 2** The overlap rate of two Gaussian components in a mixture is defined by:

$$OLR(G_1, G_2) = \begin{cases} 1 & \text{if } p(X) \text{ has one peak} \\ \frac{p(X_{Saddle})}{p(X_{Sub\_Max})} & \text{if } p(X) \text{ has two peaks} \end{cases} \tag{11}$$

where $X_{Saddle}$ is the saddle point of $pdf$, $p(X)$, $X_{Sub\_Max} = \arg(Sub\_Max_{X \in C}\ p(X))$ is the lower peak point of $pdf$ and $C$ represents the Ridge Curve. Depended on the **Theorem 2**, the $X_{Saddle}$ and $X_{Sub\_Max}$ are on the Ridge Curve. Figure 2 depicts different elements of this definition and illustrates a case of relatively high overlap rate where the saddle point is high compared to the lower peak.

The *OLR* describes the degree of overlap between two components (clusters). It is not the percentage of the data falling in the "overlapping region", nor it is linearly dependent on this percentage. The relation between *OLR* and parameters in the mixture will be further investigated in Sect. 4. In approximately speaking, as in the writing of this paper, we say that the two components (clusters) are (visually) well separated if the value of *OLR* is less than 0.6; they are partially overlapping, if the *OLR* belongs to (0.6, 0.8]; and they are strongly overlapping, if the *OLR* is greater than 0.8 (see also Figs. 11, 12, 13). These terms are utilized only for facilitating discussions.

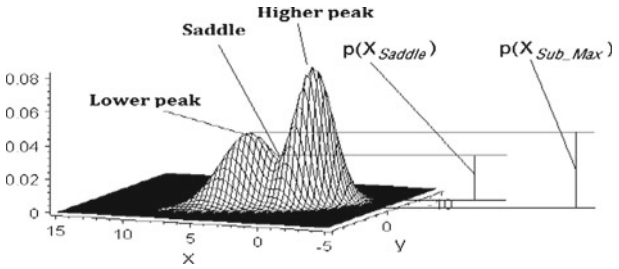


**Fig. 2** The saddle and the peaks of *pdf* used in definition of OLR

### 3.2 Algorithm for calculating *OLR*

To compute *OLR*, the stationary points of $p(X)$ need to be determined. Indeed, considering the special case of the conditions in Theorem 1, these points are the solution of the following equations:

$$\begin{cases} A_{x_1}\alpha_1 G_1 + B_{x_1}\alpha_2 G_2 = 0 \ (I) \\ A_{x_1}B_{x_2} - A_{x_2}B_{x_1} = 0 \qquad (II) \end{cases} \tag{12}$$

Expressing $x_2$ in terms of $x_1$ from (II) and substituting $x_2$ in (I), we obtain a complex equation: $M_1(x_1)e^{N_1(x_1)} + M_2(x_1)e^{N_2(x_1)} = 0$, where $M_1(x_1)$, $M_2(x_1)$, $N_1(x_1)$ and $N_2(x_1)$ are rational functions of $x_1$. This equation does not have a closed-form solution. For this reason, we have designed a numerical algorithm for calculating the stationary points of $p(X)$ based on Theorem 3. The main idea of the algorithm is to search the segment of the ridge curve $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$ 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 COLR** below computes the overlap rate of any two components (represented by $G_1$ and $G_2$) of the mixture given by Eq. 1. Using this algorithm, we can estimate the overlap rate of any two clusters in a given set of data by first estimating the mean, covariance matrix and prior probability of each cluster.

**Algorithm COLR** (for computing the *OLR* of the mixture in Eq. 1)

1. Input the parameters of two distributions ($\mu_1$, $\mu_2$, $\Sigma_1$, $\Sigma_2$, $\alpha_1$, $\alpha_2$).
2. Compute the ridge curve $A_{x_1}B_{x_2} - B_{x_1}A_{x_2} = 0$ (determining the parameters, $a$, $b$ and $c$, in Eq. 9).
3. Scan the *RC* from $\mu_1$ to $\mu_2$ by step $\delta$ (e.g. $\delta = \|\mu_1 - \mu_2\|/1000$), finding the maximum and minimum points of $p(X)$.
   (a) Let $X_0 = \mu_1$, $X_{i+1} = X_i + \delta(\mu_1 - \mu_2)$, $i = 0, 1, ..., 999$)
   (b) Compute $p(X_i)(i = 0, 1, ..., 999)$
   (c) If $p(X_i) - p(X_{i-1}) > 0$ and $p(X_i) - p(X_{i+1}) > 0$, then $X_i$ is maximum point (peak);
   (d) If $p(X_i) - p(X_{i-1}) < 0$ and $p(X_i) - p(X_{i+1}) < 0$, then $X_i$ is minimum point;
4. if $p(X)$ has one peak then the *OLR* is equal to 1, else compute the *OLR* of the two components by Eq. 11.

### 3.3 Computing OLR in the high-dimensional case

The theory on component overlap presented above can naturally be extended to the high-dimensional case. In the $d$-dimensional case ($d > 2$), the two main concepts of the theory, ridge curve and overlap rate, can be described as follows:

In the $d$-dimensional case, the overlap rate between $G_1$ and $G_2$, $OLR(G_1, G_2)$, can be defined using the same formula as (Eq. 11). $X_{Saddle}$ and $X_{Sub\_Max}$ are the stationary points of the *pdf* (Eq. 1), and satisfy

$$\begin{cases} \frac{\partial p}{\partial x_1} = A_{x_1}\alpha_1 G_1 + B_{x_1}\alpha_2 G_2 = 0 \\ \frac{\partial p}{\partial x_2} = A_{x_2}\alpha_1 G_1 + B_{x_2}\alpha_2 G_2 = 0 \\ ..... \\ \frac{\partial p}{\partial x_d} = A_{x_d}\alpha_1 G_1 + B_{x_d}\alpha_2 G_2 = 0 \end{cases} \tag{13}$$

where $A_{x_i}$ and $B_{x_i}$ ($i = 1, 2, ...d$) are defined by:

$$\begin{pmatrix} A_{x_1} \\ A_{x_2} \\ ... \\ A_{x_d} \end{pmatrix} = \nabla \| X - \mu_1 \|^2_{\Sigma_1^{-1}} = -\Sigma_1^{-1}(X - \mu_1)$$
$$\begin{pmatrix} B_{x_1} \\ B_{x_2} \\ ... \\ B_{x_d} \end{pmatrix} = \nabla \| X - \mu_2 \|^2_{\Sigma_2^{-1}} = -\Sigma_2^{-1}(X - \mu_2) \tag{14}$$

The ridge curve of the Gaussian mixture is defined by the following $d - 1$ equations:

$$\begin{cases} A_{x_1} B_{x_2} - B_{x_1} A_{x_2} = 0 \\ A_{x_2} B_{x_3} - B_{x_2} A_{x_3} = 0 \\ ..... \\ A_{x_{d-1}} B_{x_d} - B_{x_{d-1}} A_{x_d} = 0 \end{cases} \tag{15}$$

The computation of the *OLR* can be carried out by **Algorithm COLR** using Eq. 15 as the ridge curve function in Step 2.

### 3.4 Measuring the overlap of clusters in the IRIS data set

The aim of this experiment is to show how our algorithm can be used to measure degrees of overlap between user-defined clusters. We have chosen the IRIS data set because it is the most commonly used benchmark set in cluster analysis.

It is important to point out that the *OLR* theory developed in this paper presumes that the all the parameters of the mixture are given in advance. In order to use the theory to measure degrees of overlap between user-defined clusters, one has to estimate first the parameters of each component and the mixing parameters.

This problem of parameter estimation is however different from the model-based clustering (McLachlan and Basford 1988) using a mixture model because each cluster, i.e. the data for each component, is given by the user. Consequently, an independent maximum likelihood method can be used to estimate all the parameters. For instance, in the case of the IRIS data, the following formula are used to estimate their parameters:

$$\begin{cases} \mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_j \\ \Sigma_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_j - \mu_i)^T (X_j - \mu_i) \\ \quad i = 1, 2, 3 \end{cases} \tag{16}$$

Many studies have been performed on the IRIS data. Pal and Bezdek (1995) suggest that since two of the three classes overlap substantially, one can argue in favor of either 2 or 3 classes. Halgamuge and Glesner (1994) have shown that a very good classification can be obtained by using only two features. There are various degrees of overlap on the pairs of variables (features) chosen. For purposes of illustration, we will provide a precise measurement of the overlap rates between different classes as they are projected onto subspaces generated by each pair of feature components. We will also give a precise measurement of the overlap rates between different classes taken in the 4-D data.

We tested all combinations of two features. Figure 3 illustrates the IRIS data projected onto each two-feature subspace. Figure 4 shows the *pdf* s of the mixture models based on Formula 16. The overlap phenomena shown in Fig. 3 are naturally illustrated in the figures for the corresponding *pdf* s in Fig. 4. For instance, classes 2 and 3 strongly

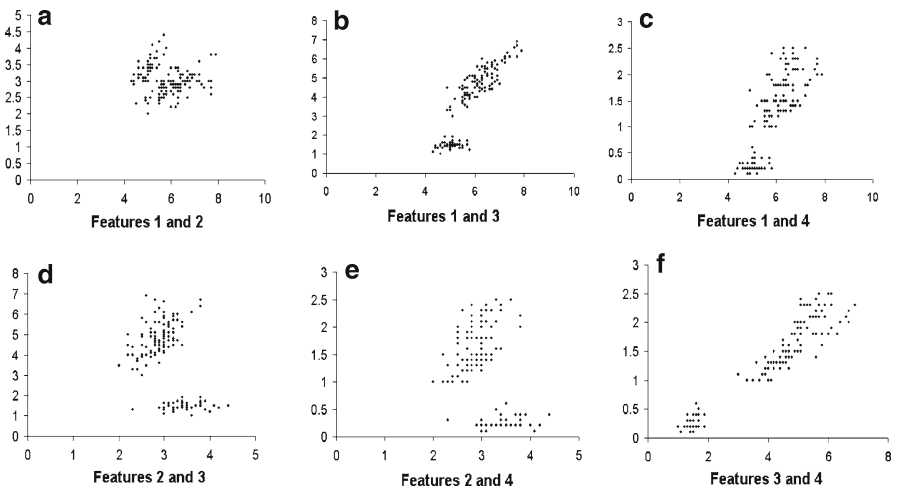


**Fig. 3** Projected IRIS data for each pair of features: **a** features (1,2), **b** features (1,3), **c** features (1,4), **d** features (2,3), **e** features (2,4) and **f** features (3,4)

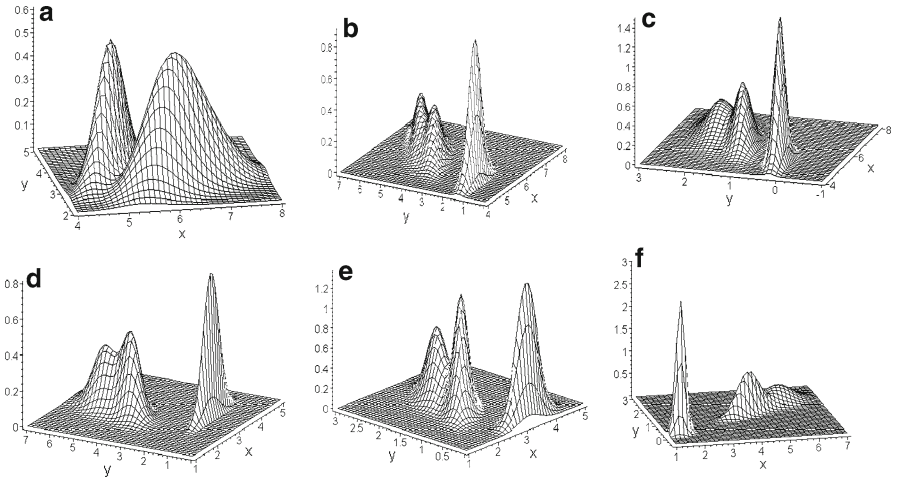


**Fig. 4** *pdf* for each pair of features: **a** features (1,2), **b** features (1,3), **c** features (1,4), **d** features (2,3), **e** features (2,4) and **f**-features (3,4)

**Table 1** The *OLR* for groups of two features in a pair of classes

| Features | (1, 2) | (1, 3) | (1, 4) | (2, 3) | (2, 4) | (3, 4) |
|---|---|---|---|---|---|---|
| Class 1, 2 | 0.115 | 0.0001 | 0.004 | 0.0001 | 0.0001 | 0.0001 |
| Class 1, 3 | 0.111 | 0 | 0 | 0 | 0 | 0 |
| Class 2, 3 | 1 | 0.683 | 0.778 | 0.895 | 0.567 | 0.776 |
| No. of classes | 2 | 2–3 | 2–3 | 2–3 | 3 | 2–3 |

overlap if features 1 and 2 are chosen (Fig. 3a). The *pdf*s of these two classes merge to form a (large) one-mode distribution in Fig. 4a. These two classes exhibit partially overlapped distributions when any other pair of feature is chosen. It should also be mentioned that class 1 is always well separated from the other classes, whichever pair of features is chosen.

**Algorithm COLR** was used to compute the *OLR* for each pair of classes. Table 1 lists the *OLR* of each pair of classes for each group of features. From this table, one can easily see that classes 2 and 3 are quite strongly overlapped for any pair of features. In particular, they are strongly overlapped if feature group (1,2) is chosen (the overlap rate between classes 2 and 3 reaches 1). The IRIS data also overlap strongly if group (2,3) is chosen. Feature group (2,4) seems to give the best overall between-class separation. We can see that **Algorithm COLR** could be used to deal with the subset feature selection problem, especially if the selection aims to improve classification accuracy.

To conclude this section, we discuss how to compute the *OLR* in the 4-D case (all other dimensional cases can be dealt with similarly). The ridge curve in this case will be determined by three equations:

**Table 2** The *OLR* for each pair of classes when all features are used

| | Class 1, 2 | Class 1, 3 | Class 2, 3 |
|---|---|---|---|
| *OLR* | $3.39 \times 10^{-6}$ | $1.66 \times 10^{-11}$ | 0.524 |

$$\begin{cases} A_{x_1} B_{x_2} - B_{x_1} A_{x_2} = 0 \\ A_{x_2} B_{x_3} - B_{x_2} A_{x_3} = 0 \\ A_{x_3} B_{x_4} - B_{x_3} A_{x_4} = 0 \end{cases} \tag{17}$$

or in other words, the projection of this curve onto any 2-D canonical subspace in Eq. 7. We used **Algorithm COLR** to compute the *OLR* for each pair of classes of the IRIS data set, with all four dimensions included. Table 2 shows the results. It is interesting to note that the overlap rate for any pair of classes is much lower than when subsets of features are used. In fact, the three classes are well separated in the original 4-D space. In the cluster analysis literature, many authors claim, erroneously, that the IRIS data could be considered as a 2-cluster data set as well as a 3-cluster data set, based solely on visual observation of the 2-D projection of the data set. However, anyone who has built a Bayes classifier with a Gaussian *pdf* for each class should have noticed that the classifier performs very well with any reasonable partition of the original data set into a training set and a test set. The overlap rate concept proposed in this paper allows for a better explanation of these results.

### 3.5 The distinction between *OLR* and classification error rate

As mentioned in Subsect. 2.2, the classification error rate can also be used to measure the degree of overlap between two components. It is at the basis of the many heuristic definitions of cluster overlap, class confusion, and class similarity (or separation) used in practice ( McLachlan and Basford 1988). In this subsection, we will use two examples to show the difference between the concept of classification error rate and the concept of overlap rate defined in this paper.

Figure 5 shows two mixtures of the same pair of components:

$$\begin{aligned} G_1(x) &= \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{x^2}{8}\right) \\ G_2(x) &= \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{8}\right) \end{aligned} \tag{18}$$

Figure 5a and c show the components of the mixtures $G = 0.5G_1(x) + 0.5G_2(x)$ and $G' = 0.7G_1(x) + 0.3G_2(x)$, while Fig. 5b and d show the *pdf*s of $G(x)$ and $G'(x)$. Figure 5a and c illustrate the area corresponding to the classification error rate, while Fig. 5b and d illustrate the overlap phenomenon that we seek to describe with *OLR*. Figure 6 illustrates the same phenomena in Fig. 5, for the 2-D case (for which the region corresponding to the most significant classification error rate cannot be illustrated as clearly as in the 1-D case). In Fig. 6,

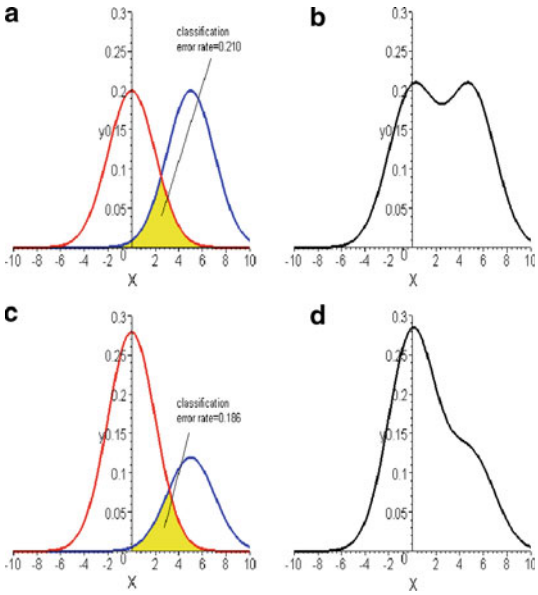


**Fig. 5** Comparison of *OLR* and Classification error rate

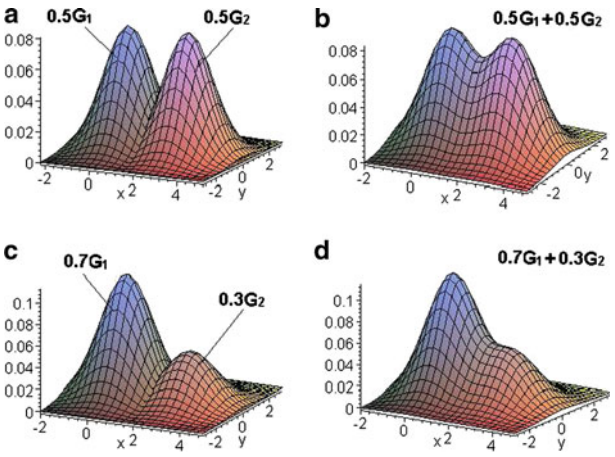


**Fig. 6** Comparison of *OLR* and Classification error rate

$$
\begin{aligned}
G_1(x, y) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \\
G_2(x, y) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}((x - 2.7)^2 + y^2)\right) \\
G(x, y) &= 0.5G_1(x, y) + 0.5G_2(x, y) \\
G'(x, y) &= 0.7G_1(x, y) + 0.3G_2(x, y)
\end{aligned} \tag{19}
$$

**Table 3** Comparison of *OLR* and classification error rate

| | 1-D case | | 2-D case | |
|---|---|---|---|---|
| Mixture | G(x) | G'(x) | G(x,y) | G'(x,y) |
| Classification error rate | 0.210 | 0.186 | 0.085 | 0.0786 |
| Overlap rate | 0.803 | 1 | 0.781 | 1 |

Table 3 summarizes the classification error rate and the *OLR* for each mixture in Figs. 5 and 6. From this table, we can remark that the classification error rate and the *OLR* do not always evolve in the same direction, contrary to the intuition that one might have. Both figures illustrate cases in which the classification error rate decreases (from $G$ to $G'$) while the overlap rate increases. Visually, we also note that the *OLR* is a better concept to measure the separation between two components. This characteristic makes the *OLR* a good alternative for measuring similarity between clusters in a hierarchical approach to clustering.

## 4 Generating simulation data sets with prescribed *OLR*s

In this section, we propose a general framework for generating truthed data sets. This is the inverse operation to computing the *OLR*: given an *OLR* value, $olr$, what are the values of the parameters a mixture must have so that the maximum of the *OLR* between each pair of components is $olr$. We restrict our discussion to the case in which there are two components in a mixture. The mixture with multiple overlapped components can be made up of various separate pairs of components. The aim of this section is to provide some ideas on how to modify the parameters in the mixture to obtain data sets with different values of the *OLR*. A more systematic use of overlap theory in generating valid data sets and evaluating the performance of clustering techniques is reported in another paper (Bouguessa et al. 2006).

### 4.1 Factors affecting the *OLR*

The problem would become much easier if the *OLR* could be expressed as an analytical function of the parameters of the mixture. In the 1-D case, Aitnouri et al. (2002) give an approximate solution to the problem based on a piece-wise linear approximation to the Gaussian components. However, this approach cannot be effectively extended to the multi-dimensional case. For this reason, we consider the factors that affect the *OLR*. The aim of the following subsections is to show the influence of different parameters of the mixture on the *OLR*. We try to show some general trends in the dependence of the *OLR* on different parameters.

Based on the discussion in the Sect. 2, we restricted our investigation to the special case defined in Theorem 1, Eq. 6. Let $G_1$ and $G_2$ be two Gaussian components of a mixture model. Without loss of generality, we suppose that the initial parameters of the two components are given by:

$$\begin{cases} G_1 : \alpha_1 = 0.5,\ \mu_1 = (0,0)^T,\ \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ G_2 : \alpha_2 = 0.5,\ \mu_2 = (3,0)^T,\ \Sigma_2 = \begin{pmatrix} 2.17 & 1.82 \\ 1.82 & 2.17 \end{pmatrix} \end{cases} \tag{20}$$

In what follows, we will show the evolution of the *OLR* when one of the parameters is varied.

#### *4.1.1 Effects of varying the mixing coefficient*

Figure 7 shows the effects of varying the mixing coefficient $\alpha_1 : 0 \rightarrow 1$. In this case, the *OLR* is a piece-wise function of $\alpha_1$. Experimental results show that the *OLR* reaches its minimum (represented by $r_{min}$) when $\alpha_1 = 0.46$. The value of $r_{min}$ depends on the components. The fact that the *OLR* reaches 1 at both ends means that the two components overlap completely when the coefficient crosses a threshold. The difference between the two covariance matrices is the main reason for the asymmetry of the *OLR* curve. Varying the value of the mixing coefficient is an easy way to obtain a mixture with an *OLR* value varying between $r_{min}$ and 1. For example, if the coefficient of $G_1$, $\alpha_1 = 0.3$, and the other parameters are same as Eq. 20, then the overlapping rate is 0.7288.

#### *4.1.2 Effects of varying the distance between the two means*

Figure 8 shows the relationship between the *OLR* and the distance between the two means. For this experiment, we varied $\mu_2$ from $(0, 0)^T$ to $(8, 0)^T$. When the two means are very close to each other (distance $< 2.16$), the *OLR* is 1. The *OLR* decreases rapidly to zero once it falls below 1 (once two partially overlapped components appear). We notice that modifying the distance between the two means leads to values of the *OLR* varying in [0, 1]. So we conclude that, for a given *OLR* ($0 \leq OLR \leq 1$), we can find a value of the distance between the two means to match the *OLR*. The distance between the two means is the most convenient parameter to control in order to obtain all possible values of the *OLR*. For example, if the $\mu_2 = (4, 0)^T$ and the other parameters keeps in Eq. 20, then the *OLR* is 0.31937.

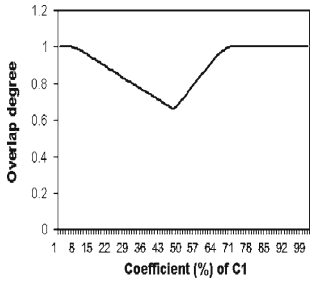


**Fig. 7** The effects of varying the mixing coefficient on the *OLR*

### 4.1.3 *Effects of varying the covariance matrix*

Describing the relationship between the *OLR* and the difference between the two covariance structures is more complex. Under the simplified assumption in Eq. 20, the contour of the *pdf* of $G_1$ is a circle, while the contour of the *pdf* of $G_2$ is an ellipse. The difference between the two covariance structures can be characterized by three factors: the angle between the two main axes of the contours (the main axis of the circle is the $x$-axis), the ratio between the principal and secondary axes of the ellipse, and the scale of the main and secondary axes. Figure 9a shows the two contours of the Gaussian mixture given by Eq. 20 and Fig. 9b shows the generated data set based on these parameters. Because the contour of $G_1$ is a circle, the angle between the two main axes is equal to the angle between the $x$-axis and the main axis of $G_2$'s contour, $\theta$, as shown in Fig. 9 (components are denoted $C_1$, $C_2$ while in Eq. 20 as $G_1$, $G_2$), and only the main and secondary axes of $G_2$'s contour are considered. We keep the mixing coefficients and means of the mixture unchanged in the following experiments.

First, we consider the effects of varying the angle between the main axes of the two contours. For this factor (see Fig. 10a), the angle $\theta$ varies from $-\pi/2$ to $\pi/2$. It is easily understood that *OLR* reaches its maximum (not necessarily 1; represented by $r_{max}$) around $\theta = 0$ and its minimum (not necessarily 0; represented by $r_{min}$) around $\theta = \pm\pi/2$ , since the main axis of the ellipse is (almost) aligned with the center of the circle (or vertical). If the given *OLR* is in $[r_{min}, r_{max}]$, we can find two angle values (positive and negative) to match the *OLR*.


**Fig. 8** The effect of varying the distance between the two means on the *OLR*

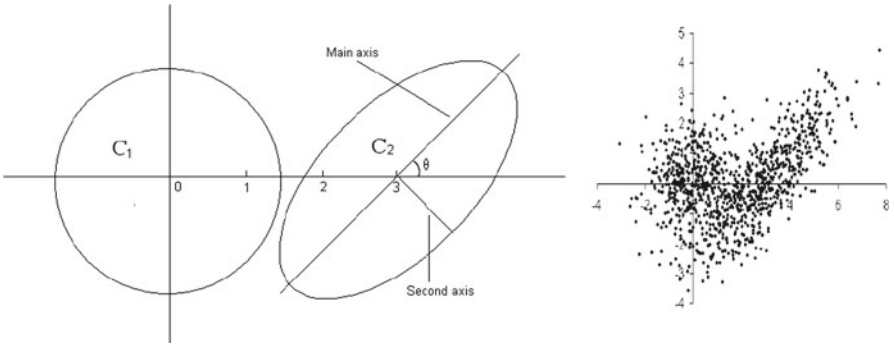


**Fig. 9** Contours of the two components and generated data set based on the parameters given by Eq. 20

Next, the effects of the ratio between the main and secondary axes are shown in Fig. 10b. The curve shows the change in the overlap rate as the secondary axis of the contour $G_2$ increases to equal the main axis. We keep the angle $\theta_2 = \pi/4$ and the length of the main axis of $G_2$ at 2.0. The length of the secondary axis varies from 0.02 to 2.0. The curve shows that the overlap rate starts from its minimum, $r_{min}$, and increases rapidly to 1.0 as the secondary axis increases. This means we can find a value of the secondary axis of $G_2$ to match a given *OLR* whose value is between $r_{min}$ and 1.

Finally, the effects of the scale of the axes are shown in Fig. 10c. The component corresponding to $G_2$ has a higher peak initially. The change in scale is performed by multiplying the lengths of both axis of the component $G_2$ by a scalar $\alpha$, which varies from 0.1 to 8. Clearly, the *OLR* increases with the scale factor $\alpha$. In this example, when $\alpha < 3.3$ the component $G_1$ remains the lower one.; when $\alpha \geq 3.3 G_2$ becomes the lower one. This explains why the *OLR* has a non-smooth change at point $P$. Finally, if the scale of both components are modified, we can obtain all *OLR* values in (0, 1].

## 4.2 Evaluation of properties of clustering algorithms using simulation data

In real applications, choosing an appropriate algorithm for cluster analysis is a big challenge, as many clustering algorithms have been proposed for different applications. Often, these algorithms are application-specific, and different clustering algorithms may yield different results on the same data set. Many studies have been done on whether particular algorithms are suitable for large and complex data sets, whether they can deal with different data types and how they handle abnormal data (Zhang and Liu 2003). These studies cannot always give an accurate indication of algorithm performance because of the lack of a measure of the difficulty of the data sets to be clustered.

In this subsection, we show how the theory of cluster overlap can be used to study properties of clustering algorithms. Specifically, we show how clustering algorithms behave in various difficult situations. These situations are described by the cluster overlap rate and the corresponding parameter values. Here, we generate simulation data sets with well-controlled overlap rates and parameters describing geometrical characteristics. Our experiments on these data sets provide a better under-

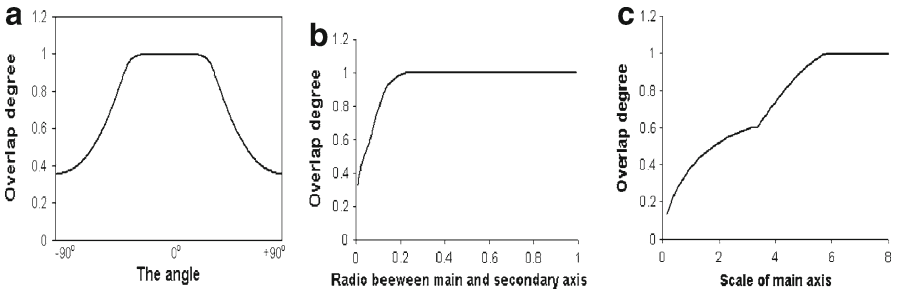


**Fig. 10** The effects of varying the two covariance matrices on the *OLR*

standing of how algorithms behave in difficult situations. For illustration purposes, we analyze three widely used algorithms: the K-means, FCM (Fuzzy C-Means) and EM (Expectation Maximization) methods, which are based on partition-based clustering, fuzzy clustering and mixture clustering, respectively. The experiments indicate that these algorithms display different properties when the clusters overlap differently. These results can aid users in their selection of a suitable algorithm in practice.

As mentioned in Sect. 4.1, the mixing coefficients (corresponding to the proportion of data in each of the two clusters), the distance between cluster means and the covariance (especially the angle between the main axes in the two components/clusters) have significant effects on the overlap rate. In the following experiment, we generate the simulation data set based on these parameters, compare the clustering results with the true clusters, and draw some conclusions about the properties of these algorithms in relation to these parameters.

### *4.2.1 Data set generation*

We generated 30 simulation data sets. Each data set contains 1,000 data points and comes from a two-component Gaussian mixture. To facilitate description of the geometric properties of the generated data sets, they are limited to two dimensions. Six of these sets are shown in the following figures.

The two data sets in Fig. 11 illustrate the difference in the overlap rate (0.97 and 0.33) caused by varying the distance between cluster centers. We generated 10 such data sets with different distance values, yielding different OLR values varying from

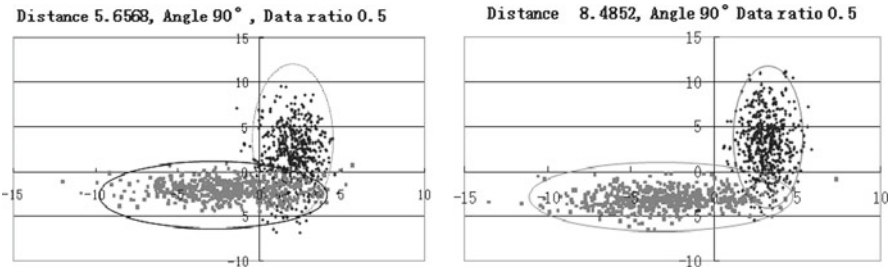


**Fig. 11** The two means change from close to far

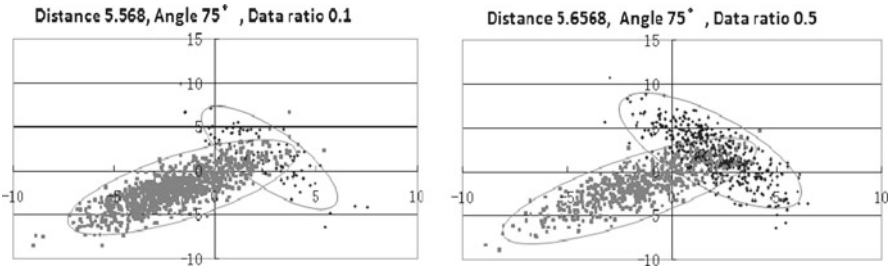


**Fig. 12** The number ratio between two clusters change from small to large

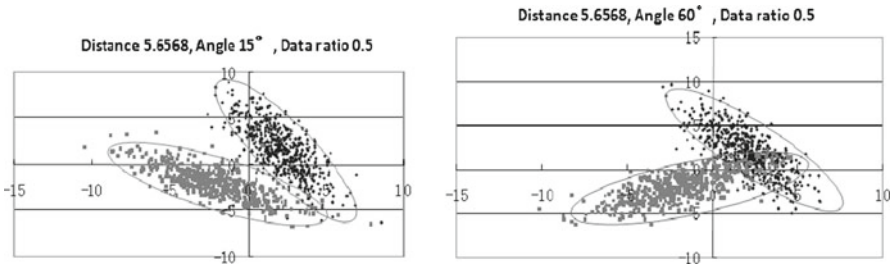


**Fig. 13** Angle of two main axis change from small to large

0.1 to 0.97. The Fig. 12 illustrates the second type of simulation data sets (the overlap rates are 1 and 0.68), which depend on the variance of the ratio between the two components. The ratio between the two components in this type varies from 0.1 to 0.5 for the 10 data sets. The corresponding OLR values vary from 1 to 0.67. Finally, Fig. 13 shows the effect of changing the angle (the overlap rates are 0.17 and 0.61). The 10 generated data sets for this third type have OLR values varying from 0.1 to 0.8.

### *4.2.2 Comparison results*

All of these data sets were clustered using the K-means, FCM and EM algorithms and the results were compared to the true clusters. The conventional clustering error rate (*CER*) defined in Eq. 21 is used as the criterion.

$$CER = \sum_{i=1}^{K} \frac{R_i}{C_i} \tag{21}$$

where $K$ is the number of clusters, $R_i$ is the number of data incorrectly clustered into the $i$th cluster and $C_i$ is the number of data in $i$th cluster. In our runs of the K-means, FCM and EM algorithms, the initial cluster centers were generated randomly. For each data set, 10 runs of each algorithm were conducted to obtain a reasonable estimate of the algorithm's performance. These results suggest certain relationships between the individual algorithms and the three parameters, which affect the overlap rate. Because of space limitations, the results are summarized in Table 4.

To understand the Table 4, it is important to remember that the *CER* always increases with increasing OLR. For each group of data sets, when cases with very small and very large OLR are present, the *CER*s of the three algorithms are similar for the data sets with the smallest and the largest OLR values. The difference between the algorithms lies in the curve of *CER* as a function of OLR. Table 4 reports comparative results in terms of the qualitative effect of OLR increase on the *CER* of each clustering algorithm.

From the Table 4, we can see that the performance of the three algorithms is sensitive to the distance between the two means. However, they behave differently in response to variation of the other two parameters. Overlap rate increases due to data ratio changes

**Table 4** Summary of experimental results on the relationships between the three clustering algorithms and different scenarios of cluster overlap (see text for more explanation of the terms used)

| Effect of OLR on clustering performance | K-means | FCM | EM |
|---|---|---|---|
| Overlap increase when the between-centers distance becomes small | Strong: CER increases increase quickly w.r.t the increase of OLR at small OLR values | Strong: similar to K-means | Strong: similar to K-means |
| Overlap increase, when the data ratio becomes large | Moderate: CER remains small when OLR value is small. CER becomes sensitive to increase of OLR when the OLR values are moderately large | Moderate effect similar to K-means. However, CER does not increase as fast for FCM as for K-means's | Weak effect: CER remains small even when OLR values are moderately large. CER becomes sensitive to increase of OLR when its values are large |
| Overlap increase when the angle becomes small | Moderate: CER remains small when OLR is small. CER becomes sensitive to increase of OLR when the OLR values are relatively large | Weak effect: CER remains small even when OLR becomes large. CER becomes sensitive to increase of OLR when the OLR values are large | Weak effect similar to FCM. However the CER does not increase as fast for EM as for FCM's |

have a moderate effect on K-means and FCM, and a weak effect on EM. Finally, overlap rate increases due to angle changes have a moderate effect on K-means, and a weak effect on FCM and EM. These results are in line with the general knowledge about the performance of the three algorithms. Detailed analysis and extension of this approach to the evaluation of clustering algorithms designed for high-dimensional data will be reported in future papers. In particular, we are investigating how to use the theory of cluster overlap to generate high-dimensional data with overlapped subspace clusters.

## 5 Discussion and conclusion

The main contribution of this paper is the establishment of a theory to explain the phenomenon of overlap in mixtures of Gaussians. The significance of this theory is that it provides a mathematically rigorous way to explain the overlap phenomenon and a computationally feasible way to calculate the degree of overlap between the

components of a mixture. It provides a foundation for generating overlapped data sets for use in validating clustering and classification algorithms.

The concept of overlap rate provides a new measure of the separability of two components. Compared to the classification error rate, the OLR provides a better explanation of the overlap phenomena. Another important property of this approach is that in addition to computing the OLR, the point in the space at which the OLR is reached is located. Apart from providing a "visual" interpretation and a better understanding of separability, this property is a key element for the inverse operation, i.e., generating simulation data with a prescribed degree of overlap.

We are currently pursuing our research in several directions. These include investigations into whether the theory can be extended to other mixtures, how to adapt the theory to explain overlapping phenomena for subspace clusters in high-dimensional spaces, and how to use the theory to develop hierarchical clustering algorithms capable of extracting clusters of arbitrary shape.

**Acknowledgments** This work has been supported by Nature Science Found of Guangdong Province, China (ID:8151503101000016), the Natural Science Foundation of Guangdong Province, P. R. China for providing a grant (No: 8351503101000001) for this project, Natural Sciences and Engineering Research Council of Canada Discovery Grant No. 121680 and Network Centers of Excellence on the Automobile of the 21st Century (NCE AUTO21) as part of a project related to navigation data construction, management and mining.

## References

Aitnouri E, Dubeau V, Wang S, Ziou D (2002) Controlling mixture component overlap for clustering algorithms evaluation. J Pattern Recog Image Anal 12(4):331–346

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

Bouguessa M, Wang S, Sun H (2006) An objective approach to cluster validation. Pattern Recogn Lett 27(13):1419–1430

Chan H, Chung A, Yu A.N.S, Wells W (2003) Clustering web content for efficient replication. In: 2003 Conference on computer vision and pattern recognition (CVPR '03), vol II

Day N (1969) Estimating the components of a mixture of two normal distributions. Biometrics 56:463–474

Do M, Vetterliyx M (2000) Texture similarity measurement using Kullback-Leibler distance on wavelet subbands. In: 2000 international conference on image processing (ICIP00), vol 3, pp 730–733

Fraley C (1998) Algorithm for model-based Gaussian hierachical clustering. SIAM J Sci Comput 20(1):270–281

Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic-Press, New York

Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 11(7):773–781

Halgamuge S, Glesner M (1994) Neural networks in designing fuzzy systems for real world applications. Fuzzy Sets and Syst 65(1):1–12

Hsu T-H (2000) An application of fuzzy clustering group-positioning analysis. Proc Natl Sci Counc ROC(C) 10:157–167

Kullback S (1959) Information theory and statistics. Wiley, New York

McLachlan G, Basford K (1988) Mixture models inference and applications to clustering. Marcel Dekker, New York

Milligan G (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrica 45(3):325–342

Nicholls K, Tudorancea C (2001) Application of fuzzy cluster analysis to Lake Simcoe crustacean zooplankton community structure. Can J Fish Aquat Sci 58(2):231–240

Pal N, Bezdek J (1995) On cluster validity for the fuzzy C-means Model. IEEE Trans Fuzzy Syst 3(3): 370–390

Ramos V, Muge F (2000) Map segmentation by colour cube genetic k-mean clustering. In: ECDL 2000, vol 1923, Lisbon, Portugal, pp 319–323

Salvi G (2003) Accent clustering in Swedish using the Bhattacharyya distance. In: 15th (ICPhS) International congress of phonetic sciences, pp 1149–1152

Sun H, Wang S (2004) Distinguishing between overlapping components in mixture models. In: Proceedings of the 2nd IASTED international conference on neural networks and computational intelligence, Switzerland, pp 102–108

Sun H, Wang S, Jiang Q (2004) FCM-based model selection algorithm for determining the number of clusters. Pattern Recogn 37(10):2027–2037

Tabbone S (1994) Edge detection, subpixel and junctions using multiple scales. PhD thesis, Institut National Polytechnique de Lorraine, France (in French)

Zhang H, Liu X (2003) The comparison of clustering methods in data mining. Comput Appl Soft 2:7–8