ISSN 2007-9737

Controlling 2D Artificial Data Mixtures Overlap

Mohammed Ouali[1,2], Walid Mahdi[1], Radhwane Gharbaoui[3], Seyyid Ahmed Medjahed[3]

[1] College of Computers and Information Technology,
Taif University,
Arabia Saudí

[2] Thales Canada Inc.,
Canada

[3] Université des Sciences et Technologie,
Département d'Informatique,
Algérie

{mouali, wmahdi}@tu.edu.sa, {radouane.gharbaoui, sa.medjahed}@univ-usto.dz

Abstract. Clustering methods are used for identifying groups of similar objects considered as homogenous set. Unfortunately, analytic performance evaluation of clustering methods is a difficult task because of their ad-hoc nature. In this paper, we propose a new test case generator of artificial data for 2 dimensional Gaussian mixtures. The proposed generator has two interesting advantages: the first one is its ability to produce simulated mixture for any number of components, while the second one resides in the fact that it formally quantifies the overlap rate which allows us to add some complexity to the data. Clustering algorithms and validity indices behavior is also analyzed by changing the overlap rate between clusters.

Keywords. Clustering algorithms, unsupervised learning, Gaussian mixture, Gaussian components overlap.

1 Introduction

Clustering methods are defined as unsupervised learning processes used to divide a set of observations into clusters [36, 38, 31, 28]. Clusters are groups of similar observations which are sufficiently far from each other. Several clustering methods are defined in the literature and all share a common nature - the difficulty in their analytical evaluation [39, 3, 9, 23, 14].

The discrete frequencies of observations and the measures of similarities between entities and clusters produce many local minima which disturb the process of clustering to converge to the correct results. Therefore, one avenue is to evaluate clustering algorithms using artificially constructed data. Many authors have categorized unsupervised classification based on criteria such as similarity or dissimilarity measures, the nature of data, and the function to be optimized [18]. Based on evaluation, the two principal categories are hierarchical methods and mobile centers methods.

The ultra-metric inequality is one of the most often used methods to generate artificial data to evaluate hierarchical algorithms [10, 22]. A large number of popular methods are based on employing the mixture model and particularly the Gaussian mixtures [15, 2, 25, 26]. The mixture models must satisfy some properties that conform with the clustering methods [24]. These patterns are summarized in two criteria: the *internal cohesion* and the *external isolation*.

Internal cohesion ensures observations within the same cluster have similar properties. External isolation ensures observation from different clusters are very dissimilar. Several works in the literature considered the Gaussian distribution as design block for clustering algorithms due its

doi: 10.13053/CyS-24-3-3326

well-known properties [15, 2]. On the other hand, several approaches have been proposed to generate artificial data. Salem and Nandy [30] proposed different structures for producing artificial observations in 2D spaces. The main rule to preserve the internal cohesion of the components mixture is to introduce empty space between the clusters despite the fact that empty space is not a sufficient condition to guarantee external isolation. In the case where the mixture components have close enough centers, no clustering method has the ability to identify the components [30]. In [34, 35], well separated data is generated for 2D and 3D cases.

The two criteria characterizing the cluster structure are strongly respected. Milligan [23] developed an algorithm for generating artificial data but was only able to avoid the total overlap for the first dimension. The claim is that avoiding overlap for the first dimension allows by transitivity to avoid total overlap for the rest of the dimensions [8, 29]. Milligan's algorithm is verified by visual inspection. Kuiper and Fisher [20] and Bayne *et al.* [5] directly manipulated a variable which measured certain parameters as separability for a simple covariance matrix of normal clusters. Blashfield [6] and Edelbrock [13] used unconstrained multivariate Gaussian mixture with fairly complex covariance structures.

This allows to obtain cluster structure and the clusters are well separated [14]. Other authors have inserted noise in well separated data to add some complexity to the obtained simulated data [30, 34, 35]. Baudry *et al.* [4] proposed a verification method to estimate the Gaussian mixture model. This work clearly distinguishes cluster structures of the mixture where the components are well separated from the Gaussian mixtures in case of total overlap. In [17, 37], the authors proposed a new artificial data generator that embeds the notion of the rate of overlap for uncorrelated 2D artificial Gaussian data.

In this paper, we propose a new automatic method for generating artificial data by controlling mixture components overlap. This work tackles two main problems: the design of an artificial data generator for correlated 2D data, and the study of the behavior of clustering algorithms and their respective validity indices by varying the rate of overlap between the mixture components. We are interested in correlated data because of the growing number of applications in computer vision and image processing, as clustering is used as the core solution to solve problems such as segmentation and image matching. In these applications, correlated data that revealed useful when combined and could be used in the clustering process are pixel gray-level, local window gray-level, and local variance.

In this paper, we will show how the overlap rate is quantified and its use as the basis block in the artificial data generator. The rest of this paper is organized as follows: section 2 briefly presents the Gaussian mixture; section 3 deals with components separation; section 4 presents the quantification of component overlap; in section 5, the control of overlap is developed; the generation algorithm and the experimental results are shown in sections 6 and 7 respectively. Finally, the conclusion is drawn with some perspectives.

2 Bivariate Correlated Gaussian Mixture Model

Mixture models are widely used in many applications because many real and natural phenomena as well as sets of data in many disciplines are based on such distributions [14, 1, 3, 25, 30]. A mixture of M Gaussian 2D components is given by:

$$P(x, y) = \sum_{j=1}^{M} \kappa_j G_j(x, y, \theta_j),$$

where $\sum_{j=1}^{M} \kappa_j = 1$ and $\theta_i = (\mu_{xj}, \mu_{yj}, \sigma_{xj}, \sigma_{yj}, \rho_j)$ denotes the parameters of the j^{th} distribution G_j. G_j is given by:

$$G_j(x, y) = A\, exp\left(-\frac{1}{2(1-\rho_j^2)}\left[\frac{t_1^2}{\sigma_{xj}^2} + \frac{t_2^2}{\sigma_{yj}^2} - \frac{2\rho_j t_1 t_2}{\sigma_{xj}\sigma_{yj}}\right]\right),$$

where $A = \frac{1}{2\pi\sigma_{xj}\sigma_{yj}\sqrt{1-\rho_j^2}}$, is real and strictly positive. $t_1 = (x - \mu_{xj})$ and $t_2 = (y - \mu_{yj})$. μ_{xj}, μ_{yj} are the component center coordinates. σ_{xj} and σ_{yj} are the standard deviations of the first and second dimension respectively. ρ_j is the correlation coefficient between the two dimensions X and Y.

3 Well Separated Components

Initially, the clustering methods and the validity indices were evaluated by using well separated data before using any other simulated data. Most works are not based on a formal way to generate artificial data and the main technique to construct isolated mixture components is visual inspection [35, 34, 11, 4].

The objective is to propose a definition that helps qualify and quantify well separated components by involving all the mixture parameters. Mixture components are considered well separated if they exhibit a minimum overlap between clusters [25, 26]; we define:

$$\begin{cases} x_{int} = \mu_1 + 4\sigma_1, \\ x_{int} = \mu_2 - 4\sigma_2, \end{cases} \tag{1}$$

where x_{int} is not really the intersection point, but for a value sufficiently far from the centers of the two components, x_{int} is approximated to be the intersection point. To be more precise in our description, x_{int} is the unique intersection point between C_1 and C_2 where C_1 (respectively C_2) is the projection of the intersection point between component Γ_1 (respectively Γ_2) and the line Δ_1 : $y = \frac{\kappa_1}{\sqrt{2\pi}\sigma_1}e^{-8}$ ($\Delta_2 : y = \frac{\kappa_2}{\sqrt{2\pi}\sigma_2}e^{-8}$). But, for x_{int} value sufficiently far from the center of the two components, x_{int} is approximated to take the form of the equation 1. In a Gaussian cluster, 99.7% of the observations belong to the interval $]\mu - 3\sigma, \mu + 3\sigma[$, which indicates that the above minimum definition implies the presence of empty spaces between data.

In [17], we presented in 2D the minimum overlap between two components Γ_1 and Γ_2 so that the intersection point after the projection satisfies:

$$\begin{cases} \Gamma_1(x_{int}, y_{int}) = \frac{\kappa_1}{2\pi\sigma_{1x}\sigma_{1y}\sqrt{1-\rho_1^2}}e^{-8}, \\ \Gamma_2(x_{int}, y_{int}) = \frac{\kappa_2}{2\pi\sigma_{2x}\sigma_{2y}\sqrt{1-\rho_2^2}}e^{-8}. \end{cases} \tag{2}$$

As in the 1D case, more than 99.7% of the observations are located inside the ellipse defined by the intersection of the plane defined by $z_1 = \frac{\kappa_1}{2\pi\sigma_{1x}\sigma_{1y}\sqrt{1-\rho_1^2}}e^{-8}$ (respectively $z_2 = \frac{\kappa_2}{2\pi\sigma_{2x}\sigma_{2y}\sqrt{1-\rho_2^2}}e^{-8}$) and Γ_1 (respectively Γ_2), where the condition of minimum overlap for the 2D data guarantees the presence of empty space between the data.

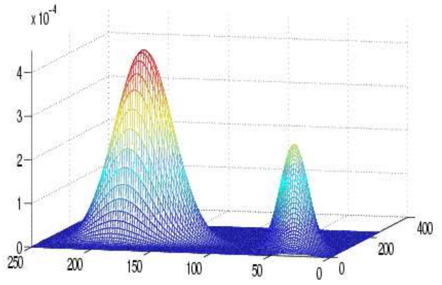


Fig. 1. An example illustrates the minimal overlap between two components of a mixture. Γ_1(0.4, 60, 40, 23, 12, 0.3); Γ_2(0.6, 73.96, 167.98, 10, 20, -0.4)

The probability density function of the generated data (*pdf*) is constrained to have the same configuration for the well separated components so that:

Definition 1: Two adjacent Gaussian components $\Gamma_1(\kappa_1,\ \mu_{x1},\ \mu_{y1},\ \sigma_{x1},\ \sigma_{y1},\ \rho_1)$ and $\Gamma_2(\kappa_2,\ \mu_{x2},\ \mu_{y2},\ \sigma_{x2},\ \sigma_{y2},\ \rho_2)$ are well separated if the intersection point between C_1 and C_2 is a unique point, where C_1 is the projection of the intersection points between Γ_1 and the plane $T_1 : z = \frac{\kappa_1}{2\pi\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}e^{-8}$, and C_2 is the projection of the intersection points between Γ_2 and the plane $T_2 : z = \frac{\kappa_2}{2\pi\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}e^{-8}$.

Formally, Γ_1 and Γ_2 are well separated if:

$$\begin{cases} \Gamma_1(x_{int}, y_{int}) = \frac{\kappa_1}{2\pi\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}e^{-8}, \\ \Gamma_2(x_{int}, y_{int}) = \frac{\kappa_2}{2\pi\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}e^{-8}, \end{cases} \tag{3}$$

where (x_{int}, y_{int}) is the coordinate of the highest intersection point from among the infinity of intersection points between the two components. Figure 1 shows a mixture of two well separated components.

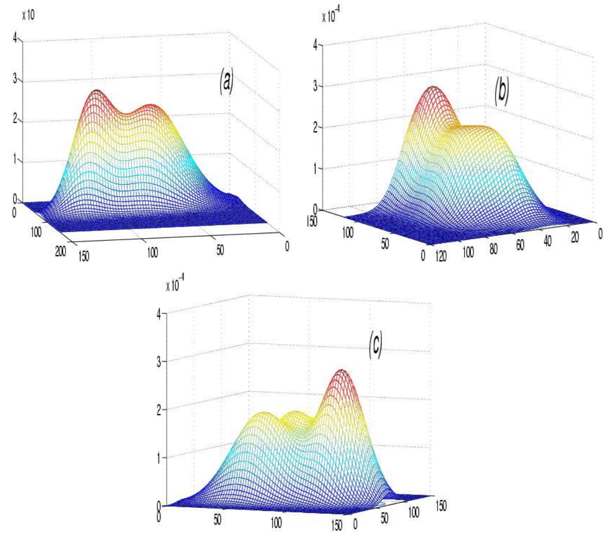


Fig. 2. Overlap Between Three Components of the Mixture in the Three Cases. (a): Total overlap between Γ_1(0.3, 60, 40, 23, 12, 0.3), Γ_2(0.3, 72, 64, 23, 12, 0.2) and Γ_3 (0.4, 110, 55, 15, 15, 0.5); (b): Maximum overlap between Γ_1(0.3, 60, 40, 23, 12, 0.3), Γ_2(0.3, 72.69, 64.15, 23, 12, 0.2) and Γ_3 (0.4, 110.15, 58.13, 15, 15, 0.5); (c): Partial overlap between Γ_1(0.3, 60, 40, 23, 12, 0.3), Γ_2(0.3, 74.59, 69.27, 23, 12, 0.2) and Γ_3 (0.4, 117.58, 63.63, 15, 15, 0.5)

3.1 Components Overlap

During the generation of a large set of data, it is important to ensure that the generated mixture components are not in a case of total overlap. Components in a case of total overlap violate the two criteria of internal cohesion and external isolation.

To better explain the meaning of total overlap, let us examine the example of figure 2. In figure 2 (a), the mixture is composed of three components; however, only two are visible; the total overlap can only be detected by visual inspection. A case of maximum overlap is shown in figure 2 (b). We can still distinguish that there are three components. In figure 2 (c), there is a partial overlap between the three components of the mixture. It is clear that the mixture is composed of three components:

It is meaningless to evaluate clustering algorithms on total overlapped structures.

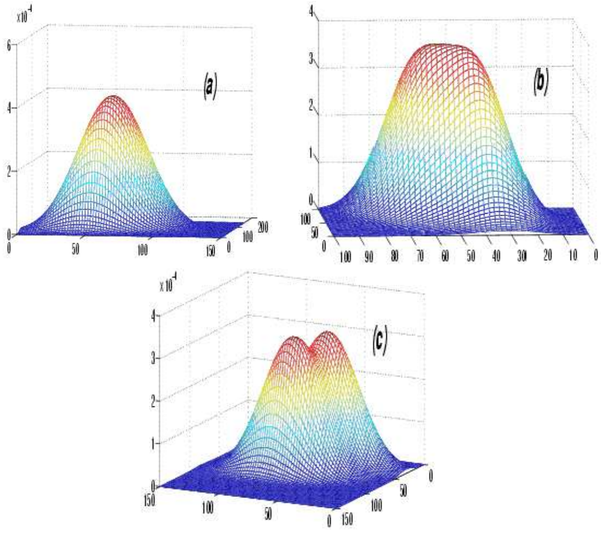


Fig. 3. Generalization and illustration of the condition (3) for two mixture equivalent components. (a) Total overlap: Γ_1 (0.5, 60, 40, 23, 12, 0.3) and Γ_2 (0.5, 72, 62, 23, 12, 0.3); (b) Maximum overlap: Γ_1 (0.5, 60, 40, 23, 12, 0.3) and Γ_2 (0.5, 74.86, 64, 23, 12, 0.3); (c) Partial overlap: Γ_1 (0.5, 60, 40, 23, 12, 0.3) and Γ_2 (0.5, 78, 66, 23, 12, 0.3)

Two components in a case of total overlap indicate that these two components form a unique component having different distribution parameters, hence it important to avoid this case.

3.2 Overlap Between Two Equivalent Bivariate Gaussian Components

In order to control components overlap, a formal quantification is needed.

In a bivariate space, let us consider two components $\Gamma_1(\mu_{x1},\ \mu_{y1},\ \sigma_x,\ \sigma_y,\ \rho_1)$ and $\Gamma_2(\mu_{x2},\ \mu_{y2},\ \sigma_x,\ \sigma_y,\ \rho_2)$, where (μ_{x1}, μ_{y1}) and (μ_{x2}, μ_{y2}) represent the centers of the first and the second components. σ_x and σ_y represent the standard deviations along each axis; ρ_1 and ρ_2 are the correlation coefficients and satisfy the equality $|\rho_1| = |\rho_2|$.

For two equivalent components, we propose the following condition for the maximum overlap:

$$\Gamma_1(x_{int}, y_{int}) = \Gamma_2(.,.) = \frac{0.5}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-1/2}, \tag{4}$$

where (x_{int}, y_{int}) represents the coordinate of the highest intersection point that has the highest value. Figure 3 illustrates the three situations related to our condition. In figure 3 (a), the value at the intersection point is higher than the value of the condition given in equation (4), which is indicative of a case of total overlap; it is impossible to visually distinguish between the two mixture components. In figure 3 (b), the intersection point obeys the condition (4). We consider this situation as the case of maximum overlap or a limit case between the total and the partial overlap. In figure 3 (c), it is clear the mixture consists of two components. The two components are in partial overlap.

The value at the intersection point is lower than the value given in condition (4). This results form a relationship between the visual inspection and the formal quantification. We will propose in the next section a definition characterizing the overlap cases. In the rest of this paper, we will use the notation $\Gamma_i(\kappa_i, \mu_{xi}, \mu_{yi}, \sigma_{xi}, \sigma_{yi}, \rho_i)$ to describe the parameters of the i^{th} component Γ_i where κ_i denotes the mixture coefficient, (μ_{xi}, μ_{yi}) are the coordinates of the components' centers, σ_{xi} and σ_{yi} are the standard deviations, ρ_i is the coefficient of correlation between the two component dimensions and $S_i = \frac{\kappa_i}{2\pi\sigma_{xi}\sigma_{yi}\sqrt{1-\rho_i^2}}e^{-1/2}$.

4 Formal Quantification of the Overlap

We propose the definition of the maximum overlap. Later we formalize the degree of overlap by the notion of rate. This definition is similar to that proposed in [17] except that in our case, the definitions are more general in order to support correlated and uncorrelated data. The overlap between components must be controlled to avoid the case of total overlap. We consider the overlap only between the two adjacent components. We will exploit the results of the previous section to propose the definitions.

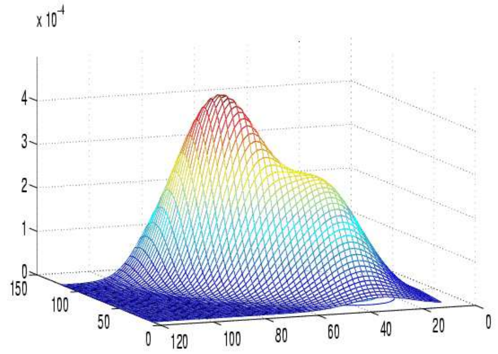


Fig. 4. Example illustrating the maximum overlap between two bivariate Gaussian mixture components, Γ_1(0.4, 60, 40, 23, 12, 0.3) and Γ_2(0.6, 88.27, 69.27, 23, 12, 0.6)

4.1 Maximum Overlap

The maximum overlap is considered as a limit between the undesirable case of total overlap and the case of partial overlap. The condition of equation (4) is extended to support non-equivalent components and we set the following definition.

Definition 2: Two adjacent Gaussian bivariate components Γ_1 and Γ_2 are in case of maximum overlap if the value at the highest intersection point $\Gamma_1(x_{int}\ ,\ y_{int}) = \min(S_1, S_2)$. Figure 4 illustrates an example of maximum overlap between two Gaussian components.

4.2 Rate of Overlap

In the literature, the notion of overlap is not quantified in a way that an artificial data can be constructed. On the other hand, there are many indices proposed to measure the shared observations or resemblance between clusters. For the most popular model, the Gaussian model, an interesting description of the fretquently used indices for computing the overlap rate between clusters is presented in [12, 32]. The Mahalanobis distance, $D_{Mah} = ((\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2))^{1/2}$, assumes that the two clusters have the same covariance matrix and the same mixture coefficients [16]. The Bhattacharyya distance is an extension of the Mahalanobis distance, $D_{Bhatt} = \frac{1}{8}(\mu_1 - \mu_2)^T[\frac{\Sigma_1+\Sigma_2}{2}]^{-1}(\mu_1 -$

$\mu_2) + \frac{1}{2}\ln\frac{|\Sigma_1+\Sigma_2|}{\sqrt{|\Sigma_1||\Sigma_2|}}$ [7]. It is difficult to use such an index because of its computing complexity, so it replaced by its upper bound $B_{Bhatt} = \sqrt{\alpha_1\alpha_2}e^{-D_{Bhatt}}$ in practical applications [16]. Other measures use the *PDF* to extract a measure for the overlap and similarity between clusters, for example the Kullback-Leibler distance $D_{kl} = p_1(x)\ln(\frac{p_1(x)}{p_2(x)}dx)$ [21]. The major inconvenience of this kind of index is that it is not symmetric. The proximity measures presented are relative and assume some conditions on the data which are in most cases simply not verified (like the equality of the components' coefficients or matrix covariance).

We propose the definition of overlap rate λ by modeling the partial overlap. This concept is based on the following points:

- The rate of overlap takes values between 0 and 1, so that the value of 1 implies the presence of maximum overlap and the value of 0 implies that the two components are "well separated".
- The overlap rate must include all the parameters of the two components: the mixture coefficients, the centers, the standard deviations and the coefficients of correlation.

Definition 3: The rate of overlap between two adjacent bivariate Gaussian components is defined as the ratio of the value at the highest intersection point to the value at the highest intersection point in the case of maximum overlap. Formally, the rate of overlap can be written as:

$$\lambda = \frac{\min(S_1, S_2)}{\min(S_{1max}, S_{2max})}.$$

These three definitions are very interesting because they employ visual inspection as a basis for the generation and verification of artificial data. Additionally, the rate of overlap definition involves symmetrically all the parameters of the two adjacent Gaussian components. We propose an algorithm for generating artificial data in order to avoid the case of total overlap and control the overlap rate λ. The parameters of the initial component are randomly generated and the parameters of the second component are computed in accordance with one of the three definitions depending on which case is to be reproduced.

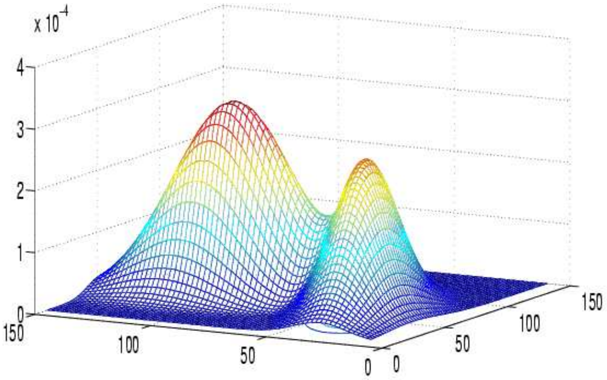


Fig. 5. Partial overlap between two bivariate Gaussian components Γ_1 : (0.4, 60, 40, 23, 12, 0.3) and Γ_2 : (0.6, 63.67, 98.13, 10, 20, -0.4) with $\lambda = 0.5$.

5 Controlling Mixture Overlap

As mentioned above, we randomly generate the parameters of the first component - the mixture coefficients, the standard deviations and the coefficients of correlation with the other components. We also introduce the angles of intersection between the components randomly. The angles of intersection are used to measure the deviation of the intersection points from the x axis. After that, we fix the centers of the components, one at a time, according to the rate of overlap.

5.1 Fixing Partial Overlap Rate

For two components $\Gamma_1(\kappa_1, \mu_{x1}, \mu_{y1}, \sigma_{x1}, \sigma_{y1}, \rho_1)$ and $\Gamma_2(\kappa_2, \mu_{x2}, \mu_{y2}, \sigma_{x2}, \sigma_{y2}, \rho_2)$, we know all the parameters of Γ_1 and $\kappa_2, \sigma_{x2}, \sigma_{y2}, \rho_2$ and we compute the center of the second component (μ_{x2}, μ_{y2}) according to the rate of overlap λ. We apply the definition of the overlap rate on the two components. There are two cases: $S_1 \geq S_2$ and $S_1 < S_2$.

Case 1: For Γ_1, after applying the overlap rate definition, we find:

$$\Gamma_1(x_{int}, y_{int}) = S_2,$$

which means that:

$$\begin{array}{l}\Gamma_1(x_{int}, y_{int}) = \\ A_1 exp\left(-\frac{1}{2(1-\rho_1^2)}\left[\frac{t_{11}^2}{\sigma_{x1}^2} + \frac{t_{12}^2}{\sigma_{y1}^2} - \frac{2\rho_1 t_{11} t_{12}}{\sigma_{x1}\sigma_{y1}}\right]\right) \\ = A_2 exp(-1/2),\end{array}$$

where: $A_i = \frac{\kappa_i}{2\pi\sigma_{xi}\sigma_{yi}\sqrt{1-p_1^2}}$ and $i \in \{1,2\}$. $t_{11} = (x_{int} - \mu_{x1})$ and $t_{12} = (y_{int} - \mu_{y1})$.

We have:

$$\begin{aligned}& a_1(x_{int} - \mu_{x1})^2 + b_1(y_{int} - \mu_{y1})^2 + \\ & c_1(x_{int} - \mu_{x1})(y_{int} - \mu_{y1}) - 1 = 0, \end{aligned} \tag{5}$$

where:

$$\begin{array}{l} e_1 = 1 - 2\ln\left(\frac{\lambda\kappa_2\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}{\kappa_1\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}\right), \\ a_1 = \frac{1}{(1-\rho_1^2)\sigma_{x1}^2 e_1}, \\ b_1 = \frac{1}{(1-\rho_1^2)\sigma_{y1}^2 e_1}, \\ c_1 = -\frac{2\rho_1}{\sigma_{x1}\sigma_{y1}(1-\rho_1^2)e_1}. \end{array} \tag{6}$$

From the inequality $S_1 \geq S_2$, we conclude that $\frac{\kappa_1}{\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1}} \geq \frac{\kappa_2}{\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2}}$. This means that $0 < \frac{\kappa_2\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}{\kappa_1\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}} \geq 1$. With $0 < \lambda \geq 1$, it is clear that $e_1 > 0$. So, we deduce that a_1 and b_1 are also strictly positive.

For the second component, by applying the same reasoning, we find:

$$\begin{aligned}& a_2(x_{int} - \mu_{x2})^2 + b_2(y_{int} - \mu_{y2})^2 + \\ & c_2(x_{int} - \mu_{x2})(y_{int} - \mu_{y2}) - 1 = 0, \end{aligned} \tag{7}$$

where:

$$\begin{cases} e_2 = 1 - 2\ln(\lambda), \\ a_2 = \frac{1}{(1-\rho_2^2)\sigma_{x2}^2 e_2}, \\ b_2 = \frac{1}{(1-\rho_2^2)\sigma_{y2}^2 e_2}, \\ c_2 = -\frac{2\rho_2}{\sigma_{x2}\sigma_{y2}(1-\rho_2^2)e_2}. \end{cases} \tag{8}$$

e_2, a_2, b_2 are also real and strictly positive.

Case 2: In this case, we find the same equations (5,7), but with these parameters for the first component:

$$\begin{array}{l} e_1 = 1 - 2\ln(\lambda), \\ a_1 = \frac{1}{(1-\rho_1^2)\sigma_{x1}^2 e_1}, \\ b_1 = \frac{1}{(1-\rho_1^2)\sigma_{y1}^2 e_1}, \\ c_1 = -\frac{2\rho_1}{\sigma_{x1}\sigma_{y1}(1-\rho_1^2)e_1}. \end{array} \tag{9}$$

and these parameters for the second component:

$$\begin{array}{l} e_2 = 1 - 2\ln\left(\frac{\lambda\kappa_1\sigma_{x2}\sigma_{y2}\sqrt{1-\rho_2^2}}{\kappa_2\sigma_{x1}\sigma_{y1}\sqrt{1-\rho_1^2}}\right), \\ a_2 = \frac{1}{(1-\rho_2^2)\sigma_{x2}^2 e_2}, \\ b_2 = \frac{1}{(1-\rho_2^2)\sigma_{y2}^2 e_2}, \\ c_2 = -\frac{2\rho_2}{\sigma_{x2}\sigma_{y2}(1-\rho_2^2)e_2}. \end{array} \tag{10}$$

In this case, a_1, b_1, e_1, a_2, b_2 and e_2 are all real and strictly positive.

In the plane defined by the equation (T) : $z = \min(S_1, S_2)$, the two equations 5 and 7 are characteristic equations of two ellipses with centers respectively at (μ_{x1}, μ_{x2}) and (μ_{x2}, μ_{x1}). This means that fixing the center of the second component fixes the center of the second ellipse. First, we compute the value of the intersection point after we compute the center of the second component. We proceed to some transformations in the referential (R), we will translate the referential after we rotate it so that the major axis of the ellipse will be parallel to the X axis of the new referential.

We proceed to translate the referential (R) by the vector $\vec{m}(\mu_{x1}, \mu_{y1})$. Equation (5) becomes:

$$a_1 x_{int}^2 + b_1 y_{int}^2 + c_1 x_{int} y_{int} - 1 = 0. \tag{11}$$

We obtain an ellipse which center is the center of referential. After translating the referential, we proceed to the rotation in which the major axis of the ellipse will be parallel to the X axis - let us call this new referential R_1. Figure 6 illustrates the referential and the angles used for the rotation. We consider the rotation angle ϕ_1. Rotation in a bivariate space is given by:

$$\begin{cases} x' = x\cos(\phi_1) - y\sin(\phi_1), \\ y' = x\sin(\phi_1) + y\cos(\phi_1), \end{cases} \tag{12}$$

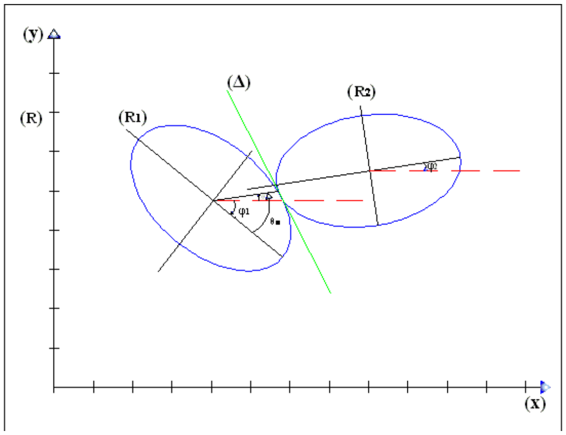


Fig. 6. Illustration of the ellipses' intersection, the different references and the angles used for the rotation

where x', y' are the coordinates in the new referential. To have a referential in which the major axis of the ellipse is parallel to the x axis, the characteristic equation of the ellipse must have the following form in the new referential:

$$a'_1 x' + b'_1 y' - 1 = 0, \tag{13}$$

where a'_1 and b'_1 are real and strictly positive because the rotation function is isometric: it preserves the distance which means that the ellipse stays an ellipse after rotation.

From equations 11, 12, and 13 and after some transformations, we have:

$$\begin{aligned} \phi_1 &= 0.5 \arctan(\tfrac{c_1}{b_1 - c_1}), \\ b'_1 &= \tfrac{b_1 \cos^2(\phi_1) - a_1 \sin^2(\phi_1)}{\cos(2\phi_1)}, \\ a'_1 &= \tfrac{a_1 \cos^2(\phi_1) - b_1 \sin^2(\phi_1)}{\cos(2\phi_1)}, \end{aligned} \tag{14}$$

where the angle θ, chosen by the user, represents the deviation of the intersection point P_{int} from the X axis. The intersection angle θ_{int} is used for two reasons. The first one is to fix one solution for computing the center of the second component as there are an infinity of solutions in which the intersection point satisfies the condition imposed by the overlap rate; the second reason is to avoid the ternary overlap between three components. The intersection angle in the new referential (R_1), after the rotation, is given by:

$$\begin{cases} \theta_{int} = \phi_1 + \theta, \\ \text{if } \theta_{int} \geq \pi/2 \text{ then } \theta_{int} = 4\pi/10, \\ \text{if } \theta_{int} \leq -\pi/2 \text{ then } \theta_{int} = -4\pi/10. \end{cases} \tag{15}$$

To avoid the ternary overlap, P_{int} is treated in the interval $]-\pi/2, \pi/2[$. For this reason, we add the two conditions cited in the equation (15) and we choose the interval $[-4\pi/10, 4\pi/10]$ as a limit. In [17], there are no conditions on the intersection point interval because, for uncorrelated data, the mixture components have by definition their axes parallel to the referential axis and there is no need for rotation; the intersection angle θ_{int} is always within the interval $]-\pi/2, \pi/2[$. From the parametrical equation of the ellipse, P_{int} coordinates x_1 and y_1 in the referential (R_1) are given by:

$$\begin{cases} t = \arctan(\sqrt{\frac{b'_1}{a'_1}} \tan(\theta_{int})), \\ x_1 = \frac{\cos(t)}{\sqrt{a'_1}}, \\ y_1 = \frac{\sin(t)}{\sqrt{b'_1}}. \end{cases} \tag{16}$$

In order to compute the second component's center, we need the value of the obliqueness at the intersection point. The obliqueness tangent is the tangent of the angle between the line tangent at the intersection point and the X axis. We have three cases; case 1: $\theta_{int} \in]0, \pi/2[$; case 2: $\theta_{int} \in]-\pi/2, 0[$ and case 3: $\theta_{int} = 0$. For the first case, the function representing this ellipse is given by:

$$f(x) = \sqrt{\frac{1 - a'_1 x'^2}{b'_1}}.$$

The value of the tangent obliqueness δ_1 in P_{int} is:

$$\delta_1 = -\frac{a'_1 x_1}{b_1 \sqrt{1 - a'_1 x_1^2}}. \tag{17}$$

For the second case, we find that the function presenting this part of ellipse is:

$$f(x) = -\sqrt{1 - \frac{a'_1 x^2}{b'_1}},$$

and the obliqueness of the line tangent on P_{int} is:

$$\delta_1 = \frac{a'_1 x_1}{b'_1\sqrt{1 - a'_1 x_1^2}}. \quad (18)$$

For the third case, where $\theta = 0$, the value of the tangent obliqueness $\delta = \infty$. The direction vector of the tangent is parallel to the Y axis. In this situation, there is no need to compute the obliqueness tangent.

$P_{int}(x_0, y_0)$ coordinates and the obliqueness of the tangent δ_0 in (R) are computed as:

$$\begin{cases} x_0 = x_1\cos(-\phi_1) - y_1\sin(-\phi_1) + \mu_{x1}, \\ y_0 = x_1\sin(-\phi_1) + y_1\cos(-\phi_1) + \mu_{y1}, \\ \delta_0 = \frac{\sin(-\phi_1)+\delta_1\cos(-\phi_1)}{\cos(-\phi_1)-\delta_1\sin(-\phi_1)}. \end{cases} \quad (19)$$

In (R_1), the direction vector of line tangent is $\vec{v}(1, \delta_1)$. We apply the rotation function to $\vec{v}$ to obtain:

$$\begin{cases} v_x = \cos(-\phi_1) - \delta_1\sin(-\phi_1), \\ v_y = \sin(-\phi_1) + \delta_1\cos(-\phi_1), \end{cases}$$

where v_x and v_y are the coordinates of the direction vector after the rotation. The obliqueness is $\delta_0 = \frac{v_y}{v_x}$.

We proceed to compute the intersection point P_{int} and the tangent. After some transformation to the tangent line at the intersection point, we extract the coordinate of P_{int} in a new referential (R_2). The new referential (R_2) has as origin the center of the second ellipse, and its axes are parallel to the axes of this ellipse. Next, the second mixture component center (μ_{x2}, μ_{y2}) is derived in the referential (R).

The treatment of the second ellipse is identical to that of the first ellipse (result of the projection of the component onto the xy plane). The referential (R) is translated by the translation vector $\vec{v}(\mu_{x2}, \mu_{y2})$. We compute the angle ϕ_2 so that the resultant referential (R_2) has an axis X parallel to the major axis of the second ellipse. After these transformations, we have:

$$\begin{cases} \phi_2 = 0.5\arctan(\frac{c_2}{b_2 - c_2}), \\ b'_2 = \frac{b_2\cos^2(\phi_2) - a_2\sin^2(\phi_2)}{\cos(2\phi_2)}, \\ a'_2 = \frac{a_2\cos^2(\phi_2) - b_2\sin^2(\phi_2)}{\cos(2\phi_2)}, \end{cases} \quad (20)$$

where the strictly positive real numbers b'_2 and a'_2 verify that the resultant characteristic equation of the second ellipse after the rotation is:

$$b'_2 y'^2 + a'_2 x'^2 = 1.$$

The value of the line obliqueness δ_2 in (R_2) is given by:

$$\delta_2 = \begin{cases} -\frac{\cos(\phi_2)}{\sin(\phi_2)}, \text{ if } (\theta_{int} = 0), \\ \frac{\sin(\phi_2)+\delta_0\cos(\phi_2)}{\cos(\phi_2)-\delta_0\sin(\phi_2)}, \text{ otherwise} \end{cases} \quad (21)$$

For the special case where $\theta_{int} = 0$, the obliqueness tangent $\delta_0 = \infty$ (the direction vector is parallel to the referential Y axis). It is easy to compute δ_2 by rotating the direction vector $\vec{V}(1, 0)$. The intersection point is finally given by: (x_2, y_2) in (R_2):

$$\begin{cases} x_2 = -\sqrt{\frac{\delta_2^2 b}{\delta_2^2 b_2 a_2 + a_2^2}}, \\ y_2 = \sqrt{\frac{1 - a_2 x_2^2}{b_2}} \text{ if } \delta_2 < 0, \\ y_2 = -\sqrt{\frac{1 - a_2 x_2^2}{b_2}} \text{ if } \delta_2 > 0. \end{cases} \quad (22)$$

x_2 can take positive values but in order to ensure that there is no total overlap between adjacent components, we choose the negative values.

We compute the coordinates x_2 and y_2 in (R) on function of μ_{x2} and μ_{y2} by applying the inverse rotation with $-\phi_2$ and by translating with $\vec{l}(\mu_{x2}, \mu_{y2})$ afterwards. Finally, μ_{x2} and μ_{y2} are given by:

$$\begin{aligned} \mu_{x2} &= x_2\cos(-\phi_2) - y_2\sin(-\phi_2) + x_0, \\ \mu_{y2} &= x_2\sin(-\phi_2) + y_2\cos(-\phi_2) + y_0. \end{aligned} \quad (23)$$

It is possible to substitute $\lambda = 1$ in the previous development to get the maximum overlap, which is a particular case of the partial overlap. Figure 7 shows five mixture components in case of maximum overlap.

By applying the definition of well separated components to the two components Γ_1 and Γ_2, we have:

$$a_1(x_{int} - \mu_{x1})^2 + b_1(y_{int} - \mu_{y1})^2 +$$

$$c_1(x_{int} - \mu_{x1})(y_{int} - \mu_{y1}) - 1 = 0,$$

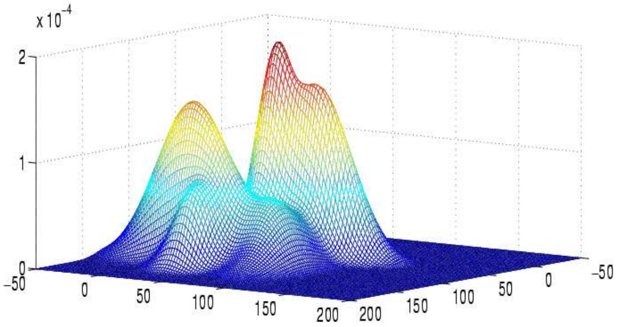


Fig. 7. Maximum overlap between five components of the mixture

Table 1. Generator initialization to obtain mixture of four components according to the variation of the overlap rate

	mixt coef	σ_{xi}	σ_{yi}	ρ_i	angle
comp 1	0.25	23	12	0.3	1.016
comp 2	0.25	25	20	0.2	0
comp 3	0.35	15	15	0.4	-1.016
comp 4	0.15	15	20	-0.2	—

Table 2. The centers obtained after the generation of bivariate artificial data

λ		0	0.5	075	1
comp. 1	μ_{x1}	60	60	60	60
	μ_{y1}	40	40	40	40
comp. 2	μ_{x2}	111.34	82.50	79.25	76.53
	μ_{y2}	167.97	93.74	85.23	77.95
comp. 3	μ_{x3}	241.93	143.40	133.05	125.63
	μ_{y3}	159.82	87.97	79.47	71.80
comp. 4	μ_{x4}	327.64	186.94	172.74	163.47
	μ_{y4}	129.27	73.46	66.5	59.89

where:

$$\begin{cases} e_1 = 16, \\ a_1 = \frac{1}{(1-\rho_1^2)\sigma_{x1}^2 e_1}, \\ b_1 = \frac{1}{(1-\rho_1^2)\sigma_{y1}^2 e_1}, \\ c_1 = -\frac{2\rho_1}{\sigma_{x1}\sigma_{y1}(1-\rho_1^2)e_1}. \end{cases} \quad (24)$$

For the second component, we find that:

$$a_2(x_{int} - \mu_{x2})^2 + b_2(y_{int} - \mu_{y2})^2 +$$

$$c_2(x_{int} - \mu_{x2})(y_{int} - \mu_{y2}) - 1 = 0, \quad (25)$$

where:

$$\begin{cases} e_2 = 1 - 0.5\ln(\lambda), \\ a_2 = \frac{1}{(1-\rho_2^2)\sigma_{x2}^2 e_2}, \\ b_2 = \frac{1}{(1-\rho_2^2)\sigma_{y2}^2 e_2}, \\ c_2 = -\frac{2\rho_2}{\sigma_{x2}\sigma_{y2}(1-\rho_2^2)e_2}. \end{cases} \quad (26)$$

The two equations 24 and 26 are characteristic equations of two ellipses in the plane $(T) : z = 0$. We follow the same equations to find the coordinates of the second component center (μ_{x2}, μ_{y2}) that satisfies components well separatedness.

6 Generation Algorithm for Gaussian Bivariate Artificial Correlated Data

In this section, the algorithm of generation of the artificial data is summarized. The general algorithm starts by introducing random values to the parameters of the first component.

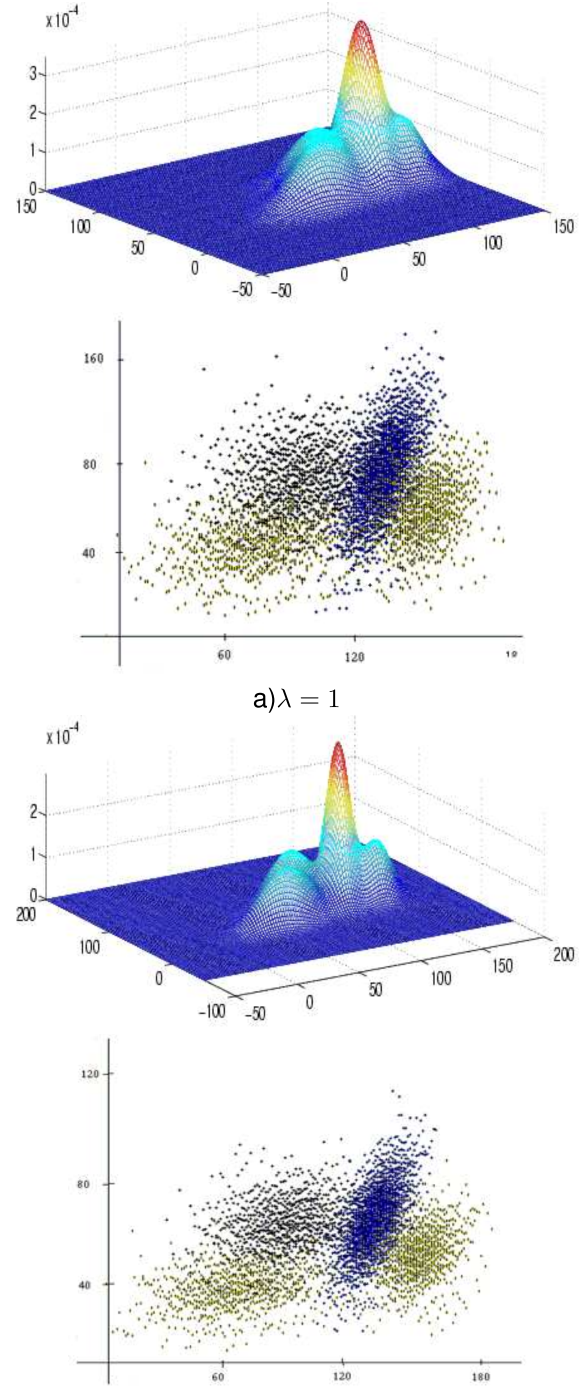
We also introduce the mixture coefficients, the standard deviations of the components, the coefficient of correlation and the deviation angles of the other components. The components' centers are derived afterwards.

In order to avoid the overlap between three components, we suggest to introduce the deviation angles θ in the interval $]-1\pi/3, 1\pi/3[$. We suggest also an interval of generation $]1, \sigma_{max}[$ for the standard deviations.

Figures 8 and 9 show a mixture of four components for each rate of overlap. If we exclude the centers of the components, the mixture components have the same parameters. Table 1 shows the generator initializations. The columns represent the mixture coefficients κ_i, the first dimension standard deviation, the second dimension standard deviation, the correlation coefficient and the intersection angle θ_{int}. θ_{int} in the table represents the angle between the current component and the next one. These parameters are obtained by varying the rate of overlap to take the values of 0, 0.5, 0.75 and 1.

We choose to give the same first component center $\mu_1 = (60, 40)$ for each of the experiments.

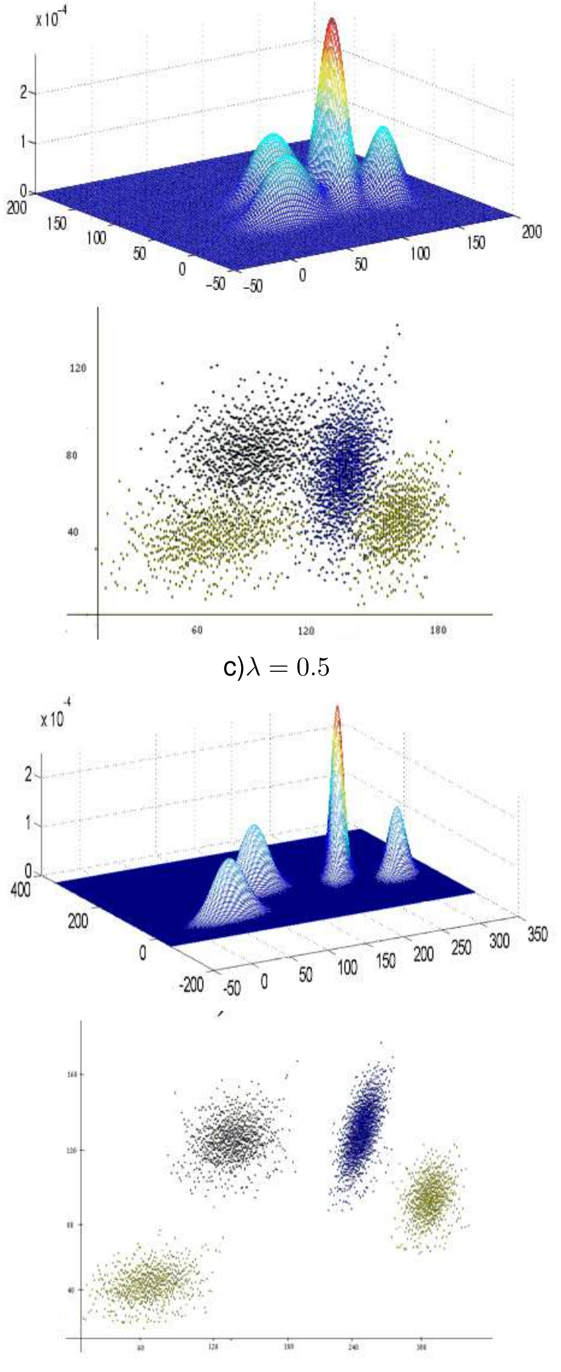
Table 2 illustrates the centers computed by the generator according to the different overlap rate values. Figures 8 and 9 show both the probability density function *pdf* and the density scatter plots. We can clearly observe that the scatter approaches each other as λ moves towards 1. It is also shown in [17] an example



a)$\lambda = 1$

b)$\lambda = 0.75$

Fig. 8. Mixture of 4 components. Overlap rate of 1 and 0.75. Density and distributions



c)$\lambda = 0.5$

d)$\lambda = 0$

Fig. 9. Mixture of 4 components. Overlap rate of 0 and 0.5. Density and distributions

Algorithm 1 Bivariate correlated Gaussian mixture

```
Number of comp.: $M$; overlap rate: $\lambda$; 1st comp. parameters: $\kappa_1$, $\mu_{x1}$ $\mu_{y1}$, $\sigma_{x1}$, $\sigma_{y1}$, $\rho_1$.
for all $i = 2, \dots, M$ do
    Generate $\kappa_i$, $\sigma_{xi}$, $\sigma_{yi}$, $\rho_i$ and $\theta_i$ / $\sum_{i=1}^{M} \kappa_i = 1$; $\theta \in [-1\pi/3; 1\pi/3]$ to inhibit the third overlap.
end for
for all $i = 2, \dots, M$ do
    if $\lambda = 0$ then
        Compute params $a_1$, $b_1$, $c_1$ and $e_1$ (24).
        Compute params $a_2$, $b_2$, $c_2$ and $e_2$ (26).
    else
        if $(\eta = \frac{\kappa_{i-1}\sigma_{xi}\sigma_{yi}\sqrt{1-\rho_i^2}}{\kappa_i\sigma_{xi-1}\sigma_{yi-1}\sqrt{1-\rho_{i-1}^2}}) \geq 1$ then
            Parameters values $a_1$, $b_1$, $c_1$, $e_1$ as in (6).
            Second comp. params as in (8).
        else
            Calculate the parameter of the first ellipse by using equation (9).
            Compute the parameters of the second ellipse from (10).
        end if
    end if
    Deduce the parameter of the second ellipse, resulting form translating and rotating the referential $R$ as in (14).
    Get the intersection angle $\theta_{int}$ from (15).
    Get the intersection point $P_{int}^{R_1}$ from (16).
    if $\theta_{int} \in ]0, \pi/2[$ then
        Use (21) to compute $\theta_1$.
    end if
    if $\theta_{int} \in ]-\pi/2, 0[$ then
        Use (22).
    end if
    $P_{int}$ and $\delta_1$ in the referential $(R)$ (19).
    Compute the second comp. parameters in referential $(R_2)$ (21).
    Compute $P_{int}^{(R_2)}$ from (22).
    Compute the center of the second comp. $(\mu_{xi}, \mu_{yi})$ as in (23).
end for
```

of four component mixture by varying λ values; however, in the current work, the example is more general, as it is based on a general algorithm and includes unconstrained Gaussian bivariate artificial correlated data.

7 Experimental Results

In this section, we present the experimental protocol and results. We propose to use k-Means, Fuzzy C-Means (FCM) and FCM-based splitting Algorithm (FBSA) [27, 19]. For validity indices, we propose to use R-Square (RS), Partition Coefficient (PC), Davies-Bouldin (DB), Xie-Benie (XB), WSJ and Classification Entropy (CE) [40].

7.1 Determination of the Number of Components

As mentioned previously, the influence of the overlap rate on clustering results is discussed. The ability of the clustering methods and the validity indices to determine the number of components is also examined.

In order to give equal opportunity to all the clustering methods to reach the correct cluster structure, we use a unique configuration for choosing the initial centers. We arrange the set of observations according to the first dimension. The k^{th} element is assumed to be the center of the k^{th} cluster so that $k = N/C + g$ where N represents the number of observations, C the number of clusters and $g = NmodC$. Tables 3 and 4 show the experimental results obtained by the proposed algorithm for 3 components and 5 components, respectively.

Tables 3 and 4 illustrate the results obtained by the proposed algorithm.

The first column represents the clustering methods used in this study: FCM, FBSA and K-Means. The second column contains the value of validity indices RS, PC, DB, XB, WSJ and CE. For each components' number, we compute the result in group according to the rate of overlap $\lambda \in \{0, 0.5, 0.75, 1\}$. Figures 8 and 9 represent the constructed data in form of clusters. As seen in Tables 3 and 4, we can easily observe that the indices RS, PC and WSJ give always the same results which means that these indices are not monotonous. For a number of observations which is sufficiently large relative to the number of components, these indices do not produce decent results.

Table 3. Results for three components

3 components						
$\lambda = 0$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	3	3	10	3
fcm	2	10	3	3	10	3
			db		rs	
k-means			3		2	
$\lambda = 0.25$						
	rs	pc	db	xb	wsj	ce
FBSA	2	10	3	3	8	3
FCM	2	10	3	3	8	3
			db		rs	
k-means			3		2	
$\lambda = 0.5$						
	rs	pc	db	xb	wsj	ce
FBSA	2	10	4	10	10	2
FCM	2	10	4	9	10	2
			db		rs	
k-means			2		2	
$\lambda = 0.75$						
	rs	pc	db	xb	wsj	ce
FBSA	2	10	2	2	10	2
FCM	2	10	2	2	10	2
			db		rs	
k-means			2		2	
$\lambda = 1$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	2	2	8	2
fcm	2	10	2	2	10	2
			db		rs	
k-means			2		2	

Table 4. Results for 5 components

5 components						
$\lambda = 0$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	5	2	10	3
fcm	2	10	5	2	10	2
			db		rs	
k-means			2		2	
$\lambda = 0.25$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	3	3	8	3
fcm	2	10	3	3	8	3
			db		rs	
k-means			3		2	
$\lambda = 0.5$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	3	3	10	3
fcm	2	10	3	3	10	3
			db		rs	
k-means			2		2	
$\lambda = 0.75$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	2	2	10	2
fcm	2	10	7	8	10	2
			db		rs	
k-means			2		2	
$\lambda = 1$						
	rs	pc	db	xb	wsj	ce
fbsa	2	10	2	2	10	2
fcm	2	10	2	2	10	2
			db		rs	
k-means			2		2	

In previous contributions [26, 17], we obtained the same result for the uni-variate and uncorrelated data. In [33], a study concerning the WSJ is presented. It is based on Bersdak suggestion, where the number of observations $N = \sqrt{C_{max}}$. The proportion N/C_{max} in that study assures good results but in our case where the number of observations $N = 3000$, it's clear that WSJ is not monotonous. For the same reason, PC and RS aren't monotonous.

We also show that by increasing the number of components or the overlap rate, the quality of the results decrease. If we look at the experimental results in Table 3, we observe that the validity indices DB, XB, PC determine the component number to be 3, but with the same overlap rate in Table 4 the above validity indices do not have the ability to identify the true number of components.

A large number of components means that there are relatively a large number of global minima to locate, so that between these global minima a large set of local minima exist where the clustering methods could wrongly converge.

From Table 3, we can easily observe that the determination of the component number becomes less and less accurate as the overlap rate increases.

From the results illustrated in Table 3, we find that all the non-monotonous validity indices are able to find the number of components when the overlap rate $\lambda = 0$; but, none of them can find the exact number of components with an overlap rate $\lambda = 1$. These results are confirmed by examining Table 4.

Contrary to the 1D experiences presented in [26], the process of clustering, in this case, cannot converge towards the true models. The curse of dimensionality affects the process for two main reasons. The first one concerns the frequency of dispersion of the data.

For the same number of observations, in 1D space, the data is distributed only on one dimension which causes the data to be more compact and the pdf appears as a continuous function with fewer local minima. However, in 2D, the data is distributed on two axes. Analytically, these spaces are viewed as local minima, and in the pdf representation, they appear as noise. The second reason concerns the overlap between more than two adjacent components.

In [26], in 1D, it is confirmed that the worst situation in which clustering methods encounter difficulty in determining the exact components number is the one where there is a component with a small deviation between two components with large standard deviations. In these circumstances, the first component overlaps beyond the second adjacent component and reaches the third component.

In 2D, there are more chances of such ternary overlap. Suppose we have three 2D components such that the intersection angle between the first and the second components is $\theta_{int} = 0.45\pi$, and the intersection angle between the second and the third components $\theta_{int} = -0.45\pi$.

In this situation, the first component is so close to the third component that they are in case of total overlap. For this reason, we limited the intersection angles to be within $]-\pi/3, \pi/3[$, despite the fact that this makes it more difficult to control in 2D.

7.2 Determination of the Clusters' Centers

In this section, we study the ability of clustering methods to determine the model parameters by knowing the number of components. The model parameters includes the mixture coefficients, the centers, the standard deviations and the correlation coefficients. The most important parameter is the centers because a small deviation from its real value has a significant influence on the other parameters.

Another point to take into consideration is that a given deviation of the components centers in the case of minimal overlap does not have the same influence in the case of maximum overlap. Because the data in maximum overlap is more compact, an error which appears negligible in minimal overlap case results in significant divergence in the mixture parameters in the maximum overlap case.

For these reasons, we have introduced a new measure ER_{avg} for computing the deviation of the mixture parameters form the real parameters. ER_{avg} is given by:

$$ER_{avg} = \frac{\sum_{i=1}^{nc} \sqrt{(\mu_{xi} - C_{xi})^2 + (\mu_{yi} - C_{yi})^2}}{nc * d_{max}},$$

where nc represents the number of clusters; μ_{xi} and μ_{yi} are the coordinates of the i^{th} component; C_{xi} and C_{yi} symbolize the i^{th} cluster coordinates; and d_{max} is the maximum distance between two components centers.

Before computing ER_{avg}, we must first associate each component center to a cluster center. There are two ways to do this. The first is to associate each component center to the nearest cluster center.

The second is to minimize the function defined as $min \sum_{\mu_i \in \mu, C_j \in C} d(\mu_i, C_i)$, where μ_i is the i^{th} component center, μ is the components centers set, and $d(a, b)$ is the Euclidean distance between a and b. Both methods produce similar results if the errors in centers are relatively small. But, in cases where the deviations in centers are large, the second method is more robust and provides better results. We use the same mixture that we used for the determination of the number of components. Table 5 illustrates the results.

Table 5. Results ER_{avg} of clustering methods

2 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.03	0.15	0.18	0.13	0.24
fcm	0.0205	0.024	0.0678	0.11	0.16
3 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.007	0.006	0.015	0.0178	0.0185
fcm	0.0025	0.0040	0.0055	0.007	0.0762
4 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.12	0.11	0.0723	0.0705	0.053
fcm	0.0812	0.00421	0.0051	0.00822	0.00842
5 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.00052	0.00481	0.0052	0.041	0.0026
fcm	0.0012	0.00551	0.015	0.0017	0.0183
6 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.054	0.0077	0.0033	0.049	0.018
fcm	0.044	0.0018	0.0241	0.035	0.032
7 clusters					
λ	0	0.25	0.5	0.75	1
k-means	0.004	0.0087	0.01	0.013	0.0048
fcm	0.004	0.00465	0.00612	0.0086	0.012

The results are proportional to the overlap rate. As the overlap rate increases, the ER_{avg} increases. In Table 5, for 5 clusters we see that $ER_{avg} = 0.0012$ when $\lambda = 0$ and $ER_{avg} = 0.0183$ when $\lambda = 1$.

A large value of λ means that there are many shared observations between data which makes the process of finding the true clusters more difficult. For the same reason, we find that increasing the number of components also increases the ER_{avg}.

8 Conclusion and Future Work

We have proposed an artificial data generator for evaluating the performance of clustering methods. The generator is used to produce artificial data for the mobile centers methods. It also benefits the hierarchical methods where the number of observations is relatively important.

Our approach is based on a formal definition and quantification of mixture components overlap. These definitions are extracted by a formal method in order to have a relationship between visual inspection of the overlap and its formal representation. We have selected three clustering algorithms to be benchmarked (FCM, FBSA and K-Means) and the validity indices RS, PC, DB, XB, WSJ and CE are used in this study.

The experiments are conducted under the same conditions including the initialization parameters and the artificial mixtures. The experimental results have shown the effectiveness and the accuracy of the produced observations especially when the overlap rate increases between components: some algorithms and validity indices outperform others and the monotonic nature of the validity indices is confirmed.

Acknowledgements

This work was funded by a Taif University Research Grant 1-437-5159. The authors are grateful to Taif University Deanship of Research for the institutional support.

References

1. **Aitnouri, E., Wang, S., & Ziou, D. (2000).** On comparison of clustering techniques for histogram pdf estimation. *Pattern Recognition Image Analysis*, Vol. 10, No. 2, pp. 206–217.
2. **Aitnouri, E., Wang, S., Ziou, D., Vaillancourt, J., & Gagnon, L. (1999).** Estimation of a multi-model's pdf using a mixture model. *Neural Parallel Scientific Computation*, Vol. 7, No. 1, pp. 103–118.
3. **Anderberg, M. (1973).** *Cluster Analysis for Applications*. New York: Academic Press.
4. **Baudry, J., Raftery, A., Celeux, G., Lo, K., & Gottardo, R. (2010).** Combining mixture component for clustering. *Journal of Computational and Graphical Statistics*, Vol. 19, No. 2, pp. 332–353.
5. **Bayne, C., Beauchamp, J., Begovich, C., & Kane, V. (1980).** Monte carlo comparisons of selected clustering procedures. *Pattern Recognition*, Vol. 12, No. 2, pp. 206–217.
6. **Blashfield, R. (1976).** Mixture model test of clusters analysis: Accurancy of four agglomerative hierarchical methods. *Psychological Bulletin*, Vol. 83, No. 3, pp. 377–388.
7. **Chen, Y., Qiu, L., Chen, W., Nguyen, L., & Katz, R. H. (2002).** Clustering web content for efficient replication. *Proceedings of the 10 IEEE International Conference on Network Protocols (ICNP'02)*, pp. 165–174.
8. **Chergas, G., Lorena, L., & Santos, R. D. (2018).** A hybrid heuristic for the overlapping cluster editing problem. *Applied Soft Computing*, Vol. 81, pp. 78–88.
9. **Cormack, R. (1971).** A review of classification. *Journal of the Royal Statistical Society*, Vol. 134, No. 3, pp. 321–367.
10. **Cunningham, K. & Ogilvie, J. (1972).** Evaluation of hierarchical grouping techniques: A preliminary stady. *The Computer Journal*, Vol. 15, pp. 209–213.
11. **Das, A. & Sil, J. (2010).** Cluter validation methods for stable cluster formation. *Canadian Journal of Artificial Intelligence, Machine Learning and pattern recognition*, Vol. 1, No. 3, pp. 26–41.
12. **Day, N. (1969).** Estimating the components of the mixture of two normal distributions. *Biometrika*, Vol. 56, No. 3, pp. 463–474.
13. **Edelbrock, C. (1979).** Comparing the accuracy of hierarchical grouping technique: The problem of classifying every body. *Multivariate Behav. Res.*, Vol. 14, No. 4, pp. 367.
14. **Everitt, B. (1974).** *Cluster Analysis*. Heinemann Educational [for] the Social Science Research Council.
15. **Everitt, B. & Hand, D. (1981).** *Finite Mixture Distribution*. London: Chapman and Holl.
16. **Fugunaga, K. (1990).** *Introduction to Pattern Recognition*. 2nd edn, Academic Press.
17. **Gharbaoui, R., Ouali, M., & Aitnouri, E. (2011).** A mixture model-based 2d data generator for performance with controlled overlap for performance evaluation. *Engineering and Technology*, Vol. 78, pp. 73–80.
18. **Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001).** Clustering validation techniques. *Intelligent Information System*.
19. **Jacques, J. & Preda, C. (2014).** Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, Vol. 71, pp. 92–106.
20. **Kuiper, F. & Fisher, L. (1975).** A monte carlo comparison between six clustring procedures. *Biometrics*, Vol. 31, No. 1.
21. **Kullback, S. (1959).** *Information Theories and Statictics*. Willey, New York.
22. **Milligan, G. (1980).** An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, Vol. 45, No. 3, pp. 325–342.
23. **Milligan, G. (1985).** An algorithm for generating artificial test clusters. *Psychometrika*, Vol. 50, No. 1, pp. 123–127.

24. **Milligan, G. & Cooper, M. (1985).** An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Vol. 50, No. 2, pp. 159–179.

25. **Ouali, M. & Aitnouri, E. (2011).** Performance evaluation of clustering technique for image segmentation. *Computer Science Journal of Maldova*, Vol. 18, No. 03, pp. 271–302.

26. **Ouali, M., Gharbaoui, R., & Aitnouri, E. (2011).** Benchmarking taxonomy for 1d clustering algorithms. *System, Signal processing and thier application (WOSSPA 2011)*, pp. 151–154.

27. **Parastar, H. & Bazrafshan, A. (2016).** Fuzzy c-means clustering for chromatographic fingerprints analysis: A gas chromatography mass spectrometry case study. *Journal of Chromatography A*, Vol. 1438, No. 1, pp. 236–243.

28. **Qiu, H., Xu, Y., Gao, L., Li, X., & Chi, L. (2016).** Multi-stage design space reduction and metamodeling optimization method based on self-organizing maps and fuzzy clustering. *Expert Systems with Applications*, Vol. 46, No. 1, pp. 180–195.

29. **Riani, M., Cerioli, A., Perrota, D., & Torti, F. (2015).** Simulating mixtures of multivariate data with fixed cluster overlap in fsda library. *Advences in Data Analysis andassification*, Vol. 9, No. 4, pp. 461–481.

30. **Salem, A. S. & Nandy, K. A. (2009).** Developpement of assessment criteria for clustering algorithms. *Pattern Analysis Application*, Vol. 12, pp. 79–98.

31. **Saltos, R. & R.Weber (2016).** A rough fuzzy approach for support vector clustering. *Information Sciences*, Vol. 339, No. 2, pp. 353–368.

32. **Sun, H. & Wang, S. (2011).** Measuring the component overlapping in the gaussian mixture model. *Data mining knowledge discovery*, Vol. 23, No. 3, pp. 479–502.

33. **Sun, H., Wang, S., & Jiang, Q. (2004).** FCM-based model selection algorithm for determinig the number of cluster. *Pattern Recognition*.

34. **Wang, W. & Zhang, Y. (2007).** On fuzzy cluter validity indices. *Fuzzy Sets and Systems*, Vol. 158, pp. 2095–2117.

35. **Wu, K. & Yang, M. (2005).** A cluster validity index for fuzzy clutering. *Pattern Recognition Letters*, Vol. 26, pp. 1275–1291.

36. **Yang, M., Chang, S., & Nataliani, Y. (2019).** Unsupervised fuzzy model-based gaussian clustering. *Information Sciences*, Vol. 481, pp. 1–23.

37. **Zhang, B., Liu, W., Zhang, H., Chen, Q., & Zhang, Z. (2016).** A note on misspecification in joint modeling of correlated data with informative cluster sizes. *Journal of Statistical Planning and Inference*, Vol. 170, No. 1, pp. 49–63.

38. **Zhao, F., Fan, J., & Liu, H. (2014).** Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local information for image segmentation. *Expert Systems with Applications*, Vol. 41, pp. 4083–4093.

39. **Zhao, K. & Lian, H. (2016).** The expectation maximization approach for bayesian quantile regression. *Computational Statistics and Data Analysis*, Vol. 96, No. 1, pp. 1–11.

40. **Zhu, E. & Ma, R. (2018).** An effective partitional clustering algorithm based on new clustering validity index. *Applied Soft Computing*, Vol. 71, pp. 608–621.

Article received on 01/01/2020; accepted on 27/02/2020.
Corresponding author is Mohammed Ouali.