Calibrated model-based evidential clustering using bootstrapping

Thierry Denoeux

Calibrated model-based evidential clustering using bootstrapping

Thierry Denœux[a,b,c]

[a] *Université de technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France*
[b] *Shanghai University, UTSEUS, Shanghai, China*
[c] *Institut universitaire de France, Paris, France*

Abstract

Evidential clustering is an approach to clustering in which cluster-membership uncertainty is represented by a collection of Dempster-Shafer mass functions forming an evidential partition. In this paper, we propose to construct these mass functions by bootstrapping finite mixture models. In the first step, we compute bootstrap percentile confidence intervals for all pairwise probabilities (the probabilities for any two objects to belong to the same class). We then construct an evidential partition such that the pairwise belief and plausibility degrees approximate the bounds of the confidence intervals. This evidential partition is calibrated, in the sense that the pairwise belief-plausibility intervals contain the true probabilities "most of the time", i.e., with a probability close to the defined confidence level. This frequentist property is verified by simulation, and the practical applicability of the method is demonstrated using several real datasets.

Keywords: Belief functions; Dempster-Shafer theory; evidence theory; resampling; unsupervised learning; mixture models.

1. Introduction

Although the first clustering algorithms were developed more than 50 years ago (see, e.g., [26] and references therein), cluster analysis is still a very active research topic today. One of the remaining open problems concerns the description and quantification of *cluster-membership uncertainty* [21, 41]. Whereas classical partitional clustering algorithms such as the c-means procedure are fully deterministic, many of the clustering algorithms used nowadays are based on ideas from fuzzy sets [4, 3, 22], possibility theory [27, 24], rough sets [31, 39] and probability theory [7, 37] to represent cluster-membership uncertainty. Recently, *evidential clustering* was introduced as a very general approach to clustering that uses the Dempster-Shafer (DS) theory of belief functions [9, 45, 19] as a model of uncertainty. At the core of the evidential clustering approach is the notion of *evidential partition* [13, 33]. Basically, an evidential partition is a vector of n mass functions $m_1, \ldots, m_n$, where

Email address: Thierry.Denoeux@utc.fr (Thierry Denœux)

n is the number of objects, and m_i is a DS mass function representing the uncertainty in the class-membership of object i [13]. Fuzzy, probabilistic, possibilistic and rough clustering are recovered as special cases corresponding to restricted forms of the mass functions [12]. Evidential clustering has been successfully applied in various domains such as machine prognosis [44], medical image processing [32, 28, 30] and analysis of social networks [52].

Different evidential clustering algorithms have been proposed to build an evidential partition of a given attribute or proximity dataset [13, 33, 14]. The EVCLUS algorithm introduced in [13] and improved in [14] consists in searching for an evidential partition such that the degrees of conflict between pairs of mass functions (m_i, m_j) match the dissimilarities d_{ij} between object pairs (i, j), up to an affine transformation. The Evidential c-Means (ECM) algorithm [33] is an alternate optimization procedure in the hard, fuzzy and possibilistic c-means family, with the difference that not only clusters, but also sets of clusters are represented by prototypes. A relational version applicable to dissimilarity data was also proposed in [34].

Evidential partitions generated by EVCLUS or ECM have been shown to be more informative than hard or fuzzy partitions. In particular, they make it possible to identify objects located in an overlapping region between two or more clusters as well as outliers, and they can easily be summarized as fuzzy or rough partitions [13, 33]. However, they are purely descriptive and unsuitable for statistical inference. In particular, if datasets are drawn repeatedly from some probability distribution, there is no guarantee that any statements derived from the evidential partitions will be true most of time.

In this paper, we propose a new method for building an evidential partition with a well-defined *frequency-calibration* property [11, 15], which can be informally described as follows. Assume that the n objects are drawn at random from some population partitioned in c classes, and each object i is described by an attribute vector $\boldsymbol{x}_i$. Given any pair of mass functions (m_i, m_j) representing uncertain information about two objects i and j, we can compute a degree of belief Bel_{ij} and a degree of plausibility Pl_{ij} that objects i and j belong to the same class [18, 29]. Now, let P_{ij} denote the *true unknown probability* that objects i and j belong to the same class, given attribute vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We will say that an evidential partition $m_1, \ldots, m_n$ is *calibrated* if, for each pair of objects i and j, the belief-plausibility interval $[Bel_{ij}, Pl_{ij}]$ is a confidence interval for the true probability P_{ij}, with some predefined confidence level $1 - \alpha$. As a consequence, the intervals $[Bel_{ij}, Pl_{ij}]$ will contain the true probability P_{ij} for a proportion at least $1 - \alpha$ of object pairs (i, j), on average.

Our approach to generate calibrated evidential partitions is based on *bootstrapping mixture models*. Model-based clustering is a flexible approach to clustering that assumes the data to be drawn from a mixture of probability distributions [1, 7, 37]. In the case of data with continuous attributes, we typically assume a Gaussian Mixture Model (GMM), in which each of the c clusters corresponds to a multivariate normal distribution [51]. The model parameters are usually estimated by the Expectation-Maximization (EM) algorithm [10, 36]. The bootstrap is a resampling technique that consists in sampling n observations from the dataset with replacement [23]. By estimating the model parameters from each bootstrap sample, we will be able to compute confidence intervals $[P_{ij}^l, P_{ij}^u]$ for each pair-

wise probability P_{ij} using the percentile method [23]. We will then compute an evidential partition $m_1, \ldots, m_n$ such that the belief-plausibility intervals $[Bel_{ij}, Pl_{ij}]$ approximate the confidence intervals $[P^l_{ij}, P^u_{ij}]$.

The rest of this paper is organized as follows. Basic definitions and results regarding evidential clustering are first recalled in Section 2. Our method is then presented in Section 3, and experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2. Evidential clustering

We first briefly introduce necessary definitions and results about DS theory in Section 2.1. The concept of evidential partition is then recalled in Section 2.2.

2.1. Dempster-Shafer theory

Let Ω be a finite set. A *mass function* on Ω is a mapping m from the power set of Ω, denoted by 2^Ω, to the interval $[0, 1]$, such that

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

Every subset A of Ω such that $m(A) > 0$ is called a *focal set* of m. When the empty set $\emptyset$ is not a focal set, m is said to be *normalized*. All mass functions will be assumed to be normalized in this paper. When all focal sets are singletons, m is said to be *Bayesian*; it is then equivalent to a probability mass functions. A mass function with only one focal set is said to be *logical*; when this focal set is a singleton, it is said to be *certain*. In DS theory, Ω represents the domain of an uncertain variable Y, and m represents evidence about Y. The mass $m(A)$ is then the degree with which the evidence supports exactly A without supporting any strict subset of A [45].

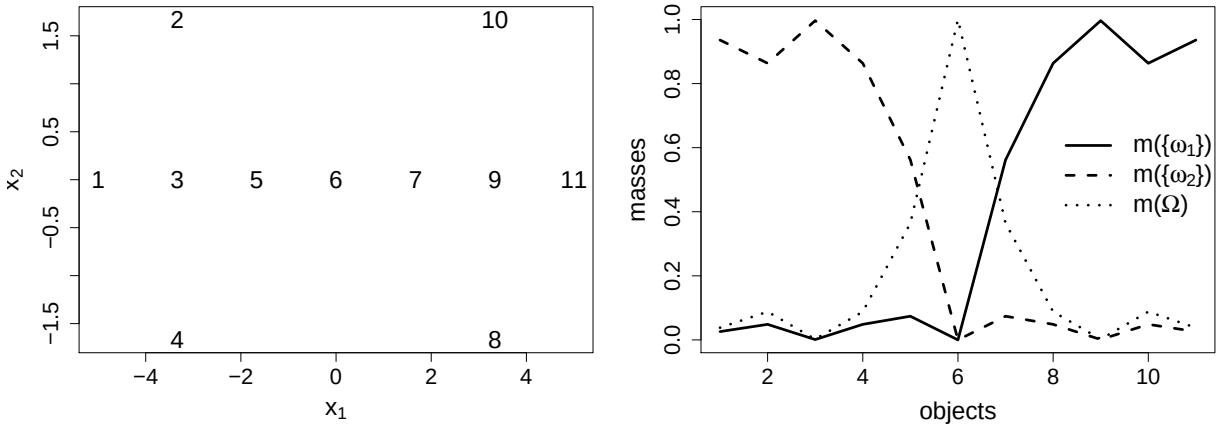
The *belief* and *plausibility* functions induced by a normalized mass function m are defined, respectively, as

$$Bel(A) := \sum_{B \subseteq A} m(B) \quad \text{and} \quad Pl(A) := \sum_{B \cap A \neq \emptyset} m(B),$$

for all $A \subseteq \Omega$. The following equalities hold: $Bel(\emptyset) = Pl(\emptyset) = 0$, $Bel(\Omega) = Pl(\Omega) = 1$, and $Pl(A) = 1 - Bel(\overline{A})$ for all $A \subseteq \Omega$, where $\overline{A}$ denotes the complement of A. The quantity $Bel(A)$ measures the total support in A, while $Pl(A)$ measures the lack of support in the complement of A. Clearly, $Bel(A) \leq Pl(A)$ for all $A \subseteq \Omega$. The three functions m, Bel and Pl are three different representations of the same information, as knowing any of them allows us to recover the other two [45].

2.2. Evidential partitions

Let $\mathcal{O}$ be a set of n objects. Each object is assumed to belong to one and only one group in $\Omega = \{\omega_1, \ldots, \omega_c\}$. An *evidential (or credal) partition* [13] is a collection $M = (m_1, \ldots, m_n)$ of n mass functions on Ω, in which m_i represents evidence about the group membership of object i. An evidential partition thus represents uncertainty about the clustering of objects in $\mathcal{O}$. The notion of evidential encompasses several classical clustering structures [12]:



(a) (b)

Figure 1: Butterfly dataset (a) and evidential partition with $c = 2$ obtained by ECM (b).

- When mass functions m_i are certain, then M is equivalent to a hard partition; this case corresponds to full certainty about the group of each object.
- When mass functions are Bayesian, then M boils down to a fuzzy partition, where the degree of membership u_{ik} of object i to group k is $u_{ik} = Bel_i(\{\omega_k\}) = Pl_i(\{\omega_k\}) \in [0, 1]$.
- When each mass function m_i is logical with focal set $A_i \subseteq \Omega$, m_i is equivalent to a rough partition [40]. The lower and upper approximations of cluster ω_k are then defined, respectively, as the set of objects that *surely* belong to group ω_k, and the set of objects that *possibly* belong to group ω_k; they are formally given by

$$\omega_k^l := \{i \in \mathcal{O} | A_i = \{\omega_k\}\} \quad \text{and} \quad \omega_k^u := \{i \in \mathcal{O} | \omega_k \in A_i\}. \tag{1}$$

  We then have $Bel_i(\{\omega_k\}) = I[i \in \omega_k^l]$ and $Pl_i(\{\omega_k\}) = I[i \in \omega_k^u]$, where $I[\cdot]$ denotes the indicator function.

Example 1. *Consider the* `Butterfly` *data displayed in Figure 1a, consisting in 11 objects described by two attributes. Figure 1b shows a normalized evidential partition of these data obtained by ECM, with $c = 2$ clusters. (An evidential partition is said to be normalized if it is composed of normalized mass functions). We can see, for instance, that object 9, which is situated in the center of the rightmost cluster ω_1, has a mass function m_9 such that $m_9(\{\omega_1\}) \approx 1$, while object 6, which is located between clusters ω_1 and ω_2, is assigned a mass function m_6 verifying $m_6(\Omega) \approx 1$ with $\Omega = \{\omega_1, \omega_2\}$.*

Given two distinct objects i and j with corresponding normalized mass functions m_i and m_j, we may consider the set $\Theta_{ij} = \{s_{ij}, \neg s_{ij}\}$, where s_{ij} denotes the proposition "Objects

i and j belong to the same cluster" and $\neg s_{ij}$ is the negation of s. As shown in [18], the normalized mass function m_{ij} on Θ_{ij} derived from m_i and m_j has the following expression:

$$m_{ij}(\{s_{ij}\}) = \sum_{k=1}^{c} m_i(\{\omega_k\}) m_j(\{\omega_k\}) \tag{2a}$$

$$m_{ij}(\{\neg s_{ij}\}) = \sum_{A \cap B = \emptyset} m_i(A) m_j(B) \tag{2b}$$

$$m_{ij}(\Theta_{ij}) = \sum_{A \cap B \neq \emptyset} m_i(A) m_j(B) - \sum_{k=1}^{c} m_i(\{\omega_k\}) m_j(\{\omega_k\}). \tag{2c}$$

Thus, the belief and plausibility that objects i and j belong to the same class are given, respectively, by

$$Bel_{ij}(\{s_{ij}\}) = m_{ij}(\{s_{ij}\}) = \sum_{k=1}^{c} m_i(\{\omega_k\}) m_j(\{\omega_k\}) \tag{3a}$$

and

$$Pl_{ij}(\{s_{ij}\}) = m_{ij}(\{s_{ij}\}) + m_{ij}(\Theta_{ij}) = \sum_{A \cap B \neq \emptyset} m_i(A) m_j(B). \tag{3b}$$

Given an evidential partition $M = (m_1, \ldots, m_2)$, the tuple $R = (m_{ij})_{1 \leq i < j \leq n}$ is called the *relational representation* of M [18].

Example 2. *Consider objects 4 and 5 the Example 1 (see Figure 1). We have*

$$m_4(\{\omega_1\}) = 0.049, \quad m_4(\{\omega_2\}) = 0.863, \quad m_4(\Omega) = 0.088$$

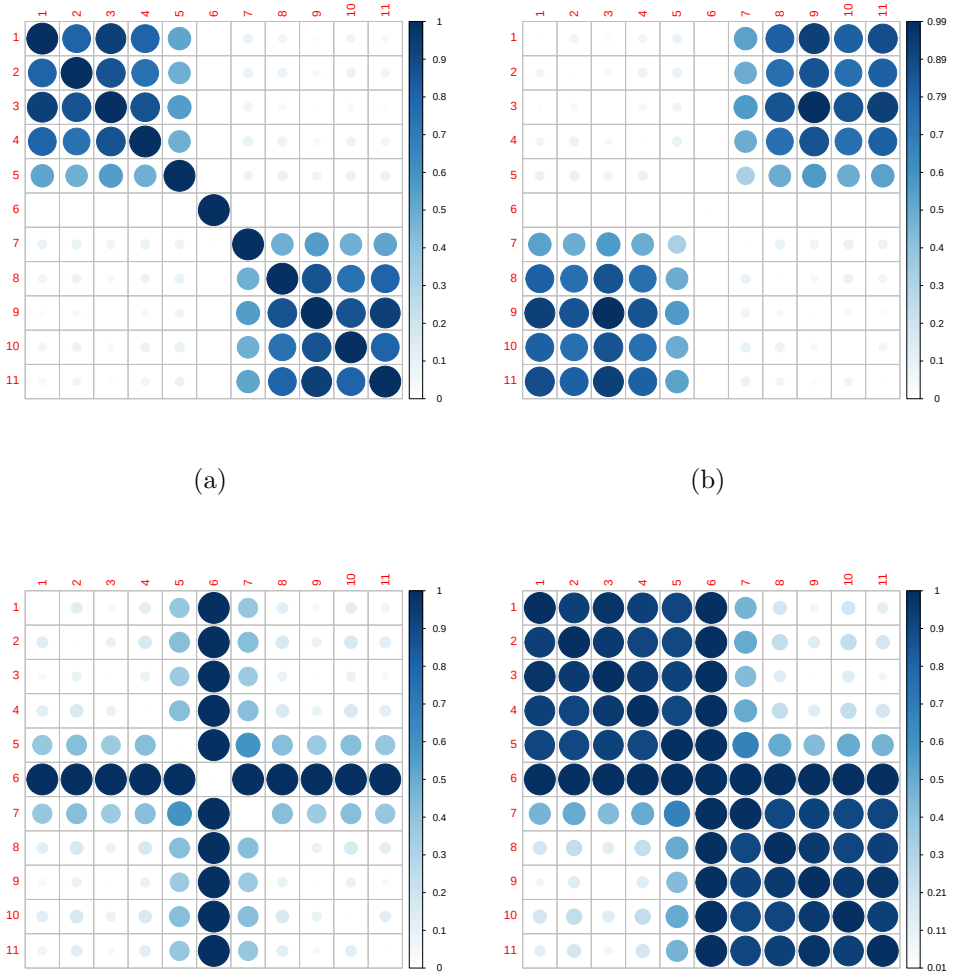
and

$$m_5(\{\omega_1\}) = 0.074, \quad m_5(\{\omega_2\}) = 0.558, \quad m_5(\Omega) = 0.368.$$

Consequently, we have

$$\begin{aligned} m_{45}(\{s_{45}\}) &= 0.049 \times 0.074 + 0.863 \times 0.558 \approx 0.485 \\ m_{45}(\{\neg s_{45}\}) &= 0.049 \times 0.558 + 0.863 \times 0.074 \approx 0.0912 \\ m_{45}(\Theta_{45}) &\approx 1 - 0.485 - 0.0912 = 0.423. \end{aligned}$$

The degree of belief that objects 4 and 5 belong to the same class is 0.485, and the degree of plausibility is $0.485 + 0.423 = 0.908$. *Figure 2 displays the complete relational representation of the evidential partition of the* `Butterfly` *data shown in Figure 1b. The matrices containing* $m_{ij}(\{s_{ij}\})$, $m_{ij}(\{\neg s_{ij}\})$ *and* $m_{ij}(\Theta_{ij})$ *are represented graphically in Figures 2a, 2b and 2c, respectively. The pairwise plausibilities* $Pl_{ij}(\{s_{ij}\})$ *are represented in Figure 2d.*



(a) (b)

(c) (d)

Figure 2: Relational representation of the evidential partition of the `Butterfly` data shown in Figure 1b: masses $m_{ij}(\{s_{ij}\})$ (a), $m_{ij}(\{\neg s_{ij}\})$ (b), $m_{ij}(\Theta_{ij})$ (c) and pairwise plausibilities $Pl_{ij}(\{s_{ij}\})$ (d).

3. Computed calibrated evidential partitions

In this section, we describe our method for quantifying the uncertainty of model-based clustering using an evidential partition with well-defined properties with respect to the unknown true partition. The assumptions will first be stated in Section 3.1. A method to compute bootstrap confidence intervals on pairwise probabilities P_{ij} will then be described in Section 3.2. Finally, an algorithm for computing an evidential partition from these confidence intervals will be introduced in Section 3.3.

3.1. Assumptions

We consider a population of objects, each one described by an attribute vector $\boldsymbol{X} \in \mathbb{R}^d$ and by a class variable $Y \in \Omega = \{1, \ldots, c\}$. The conditional distribution of $\boldsymbol{X}$ given $Y = k$ is described by a probability density function (pdf) $p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is a vector of parameters. The marginal distribution of $\boldsymbol{X}$ is, thus, a mixture distribution with pdf

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{c} \pi_k p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$$

where $\pi_k = \mathbb{P}(Y = k)$, $k = 1, \ldots, c$ are the prior class densities, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c, \pi_1, \ldots, \pi_c)$ is the vector of all parameters in the model. The conditional probability $\pi_k(\boldsymbol{x}; \boldsymbol{\theta})$ that $Y = k$ given $\boldsymbol{X} = \boldsymbol{x}$ can be computed using Bayes' theorem as

$$\pi_k(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{p_k(\boldsymbol{x}; \boldsymbol{\theta}_k)\pi_k}{\sum_{\ell=1}^{c} p_\ell(\boldsymbol{x}; \boldsymbol{\theta}_\ell)\pi_\ell}.$$

Let $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a dataset composed of n attribute vectors describing n objects. We assume that $\mathcal{D}$ is a realization of an i.i.d. sample from $\boldsymbol{X}$, and we want to quantify the uncertainty about the classes $y_1, \ldots, y_n$ of the n objects. If parameter $\boldsymbol{\theta}$ was known, then the uncertainty about y_i could be described by the conditional class probabilities $\pi_k(x_i; \boldsymbol{\theta})$, $k = 1, \ldots, c$, and the probability $P_{ij}(\boldsymbol{\theta})$ that objects i and j belong to the same class could be computed as

$$P_{ij}(\boldsymbol{\theta}) := \mathbb{P}(Y_i = Y_j \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{c} \pi_k(\boldsymbol{x}_i; \boldsymbol{\theta})\, \pi_k(\boldsymbol{x}_j; \boldsymbol{\theta}). \tag{4}$$

In usual situations, parameter $\boldsymbol{\theta}$ is unknown and it needs to be estimated from the data. Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ obtained, e.g., using the EM algorithm [10, 36]. The estimated conditional class probabilities are $\widehat{\pi}_{ik} := \pi_k(x_i; \widehat{\boldsymbol{\theta}})$, $k = 1, \ldots, c$; they constitute a fuzzy partition of the dataset. The MLE of $P_{ij}(\boldsymbol{\theta})$ is $P_{ij}(\widehat{\boldsymbol{\theta}}) := \sum_{k=1}^{c} \widehat{\pi}_{ik}\widehat{\pi}_{jk}$ for all $(i, j) \in \{1, \ldots, n\}^2$. However, these point probability estimates do not adequately reflect group-membership uncertainty, because they do not account for the uncertainty on $\boldsymbol{\theta}$. In the next section, we propose a method to compute approximate confidence intervals on pairwise probabilities $P_{ij}(\boldsymbol{\theta})$.

3.2. Confidence intervals on pairwise probabilities

Let us consider two *fixed* vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from dataset $\mathcal{D}$, and an i.i.d. random sample $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$ from $p(\boldsymbol{x}; \boldsymbol{\theta})$. A *confidence interval* on $P_{ij}(\boldsymbol{\theta})$ at level $1-\alpha$ is a random interval $\left[P^l_{ij}(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n), P^u_{ij}(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n)\right]$ such that, for all $\boldsymbol{\theta}$,

$$\mathbb{P}\left(P^l_{ij}(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n) \leq P_{ij}(\boldsymbol{\theta}) \leq P^u_{ij}(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n)\right) \geq 1 - \alpha.$$

Approximate confidence intervals on $P_{ij}(\boldsymbol{\theta})$ can be obtained in several different ways. One approach is to use the asymptotic normality of the MLE and to estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the observed information matrix. MacLachlan and Krishnan [36, Chapter 4] review different methods for computing or approximating the observed information matrix, and MacLachlan and Basford [35, Chapter 2] give an approximate analytical expression for the case of a Gaussian mixture. Estimates $\widehat{v}_{ij}$ of the variance v_{ij} of $P_{ij}(\widehat{\boldsymbol{\theta}})$ could then be obtained by the delta method, leading to the following standard confidence interval:

$$P_{ij}(\widehat{\boldsymbol{\theta}}) \pm u_{1-\alpha/2}\sqrt{\widehat{v}_{ij}}, \tag{5}$$

where $u_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution. Standard confidence intervals are consistent, but they are based on asymptotic approximations that can be quite inaccurate in practice [20]. As noted in [38], "in the case of mixture models large sample sizes are required for the asymptotics to give a reasonable approximation". In our case, the estimates $P_{ij}(\widehat{\boldsymbol{\theta}})$ take values in $[0, 1]$, and their distribution can be very asymmetric for small n, as will be shown experimentally in Example 3 below (Figure 4) and in Section 4.1 (Figure 8).

Bootstrap confidence intervals can be seen as algorithms for improving standard intervals such as (5) without human intervention [20]. Given a realization $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$ of the random sample, a nonparameteric bootstrap "pseudo-sample" is generated by drawing n observations randomly from $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$ with replacement. Repeating this operation B times, we obtain B pseudo-samples $\{\boldsymbol{x}'_{b1}, \ldots, \boldsymbol{x}'_{bn}\}_{b=1}^B$ and the corresponding estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_B$ of $\boldsymbol{\theta}$ (computed using the EM algorithm). The simplest technique for computing approximate confidence intervals using this approach is the *bootstrap percentile (BP)* method [23]. The BP confidence interval for $P_{ij}(\boldsymbol{\theta})$ is defined by the $\alpha/2$ and $1-\alpha/2$ quantiles of $P_{ij}(\widehat{\boldsymbol{\theta}}_1), \ldots, P_{ij}(\widehat{\boldsymbol{\theta}}_B)$, which will be denoted as P^l_{ij} and P^l_{ij}. Because the original dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ was generated from the same distribution as $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$, we can use it to compute bootstrap confidence intervals for any pair (i, j) of objects. The procedure is summarized in Algorithm 1. For previous applications of the bootstrap approach to model-based clustering, see [38] and references therein.

Under general conditions stated in [46, Theorem 4.1], BP confidence intervals are consistent, i.e., we have

$$\mathbb{P}\left(P^l_{ij} \leq P_{ij}(\boldsymbol{\theta}) \leq P^u_{ij}\right) \to 1 - \alpha \tag{6}$$

as $n \to \infty$. As shown by Davison and Hinkley [8, page 213], equi-tailed BP confidence intervals such as (6) are superior to standard confidence intervals such as (5), in the sense that they are second-order accurate, i.e., we have

$$\mathbb{P}\left(P^l_{ij} \leq P_{ij}(\boldsymbol{\theta}) \leq P^u_{ij}\right) = 1 - \alpha + O(n^{-1}), \tag{7}$$

whereas the coverage probability of normal approximation confidence intervals is $1-\alpha+O(n^{-1/2})$. More sophisticated procedures such as the bootstrap accelerated bias corrected (BC_a) method have also been developed to further improve the performance of BP confidence intervals [23], but these methods depend on additional coefficients that are not easy to determine. As confidence intervals on $P_{ij}(\boldsymbol{\theta})$ need to be computed for each of the $n(n-1)/2$ pairs of objects, we will stick to the simple BP method. As will be shown in Section 4.1, the confidence intervals computed by this method have coverage probabilities close to their nominal values, provided the model is correctly specified.

As a final argument in favor of the bootstrap as compared to the normal approximation method, we can observe that the latter approach relies on the calculation of the information matrix, which can very cumbersome and has to be carried out for each new model. Even if we limit ourselves to the family of Gaussian mixture models, we usually impose various restrictions on the parameters (as will be shown in Section 4.1), resulting in different expressions for the information matrix. Furthermore, with some covariance structures, we use non-differentiable orthogonal matrices, which prohibits the information matrix-based approach [38]. For non-normal models, the calculations often become intractable. In contrast, the bootstrap method can be applied without modification to any model. This advantage does come at the cost of heavier computation but, as we will see in Section 4, the computing time remains manageable on a personal computer with moderate size datasets, for which the method is useful (with large datasets, the second-order uncertainty on membership probabilities can often be neglected anyway).

Algorithm 1 Generation of BP confidence intervals on pairwise probabilities.

Require: Dataset $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$, model $p(\cdot;\boldsymbol{\theta})$, number of bootstrap samples B, confidence level $1-\alpha$

for $b=1$ **to** B **do**
  Draw $\boldsymbol{x}_{b1},\dots,\boldsymbol{x}_{bn}$ from $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ with replacement
  Compute the MLE $\widehat{\boldsymbol{\theta}}_b$ from $\boldsymbol{x}_{b1},\dots,\boldsymbol{x}_{bn}$ using the EM algorithm
  for all $i<j$ **do**
    Compute $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$
  end for
end for
for all $i<j$ **do**
  $P_{ij}^{l} := \mathsf{Quantile}\left(\left\{P_{ij}(\widehat{\boldsymbol{\theta}}_b)\right\}_{b=1}^{B}; \frac{\alpha}{2}\right)$
  $P_{ij}^{u} := \mathsf{Quantile}\left(\left\{P_{ij}(\widehat{\boldsymbol{\theta}}_b)\right\}_{b=1}^{B}; 1-\frac{\alpha}{2}\right)$
end for

Example 3. *As an example, we consider the dataset shown in Figure 3, consisting of* $n=30$ *two-dimensional vectors drawn from a mixture of* $c=3$ *components with the following parameters:*

$$\boldsymbol{\mu}_1 := (0,1)^T, \quad \boldsymbol{\mu}_2 := (1,0)^T, \quad \boldsymbol{\mu}_3 := (1,1)^T,$$

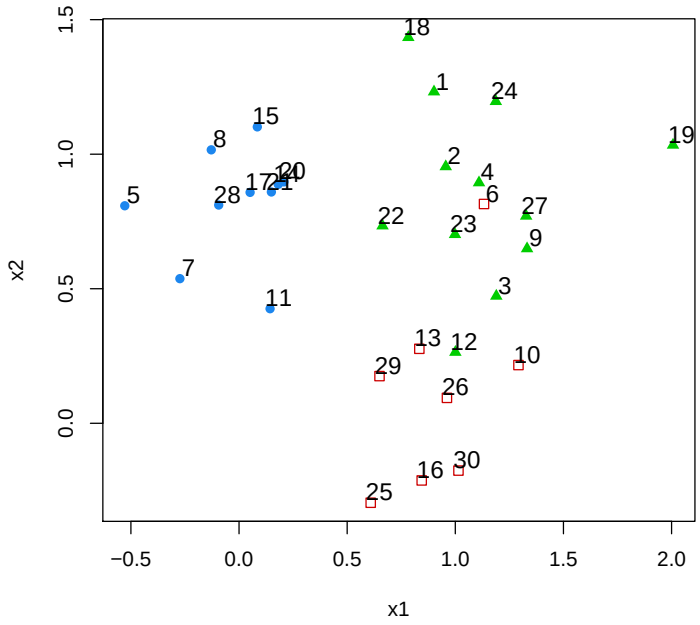


Figure 3: Dataset of Example 3.

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 := \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \quad \pi_1 = \pi_2 = \pi_3 := 1/3.$$

We applied the above method with $B = 1000$*, assuming the true model (spherical classes with equal volume). Figure 4 shows histograms of the bootstrap estimates* $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$ *and the bounds of the percentile 90% confidence interval* P_{ij}^l, P_{ij}^u *for four pairs of points. We can see that points 11 and 29 have a low probability* $P_{11,29}$ *of belonging to the same class, and the probability is well estimated with a narrow confidence interval. Point pairs (24,19) and (26,30) have a high probability of belonging to the same class, and the corresponding confidence interval is also narrow. In contrast, the true probability that points 22 and 23 belong to the same class is approximately equal to 0.7, and the corresponding confidence interval is quite large.*

3.3. Construction of an evidential partition

The $n(n-1)/2$ confidence intervals computed as described in the previous section are not easily interpretable. To obtain a simple and more user-friendly representation, we propose to construction an evidential partition $M = (m_1, \ldots, m_n)$ such that, for all pairs (i, j) of objects, $Bel_{ij}(\{s_{ij}\})$ and $Pl_{ij}(\{s_{ij}\})$ as computed by (3) approximate, respectively, the confidence bounds P_{ij}^l and P_{ij}^u. More precisely, we want to find M that minimizes the error function

$$J(M) := \sum_{i<j} \left(Bel_{ij}(\{s_{ij}\}) - P_{ij}^l\right)^2 + \left(Pl_{ij}(\{s_{ij}\}) - P_{ij}^u\right)^2. \tag{8}$$

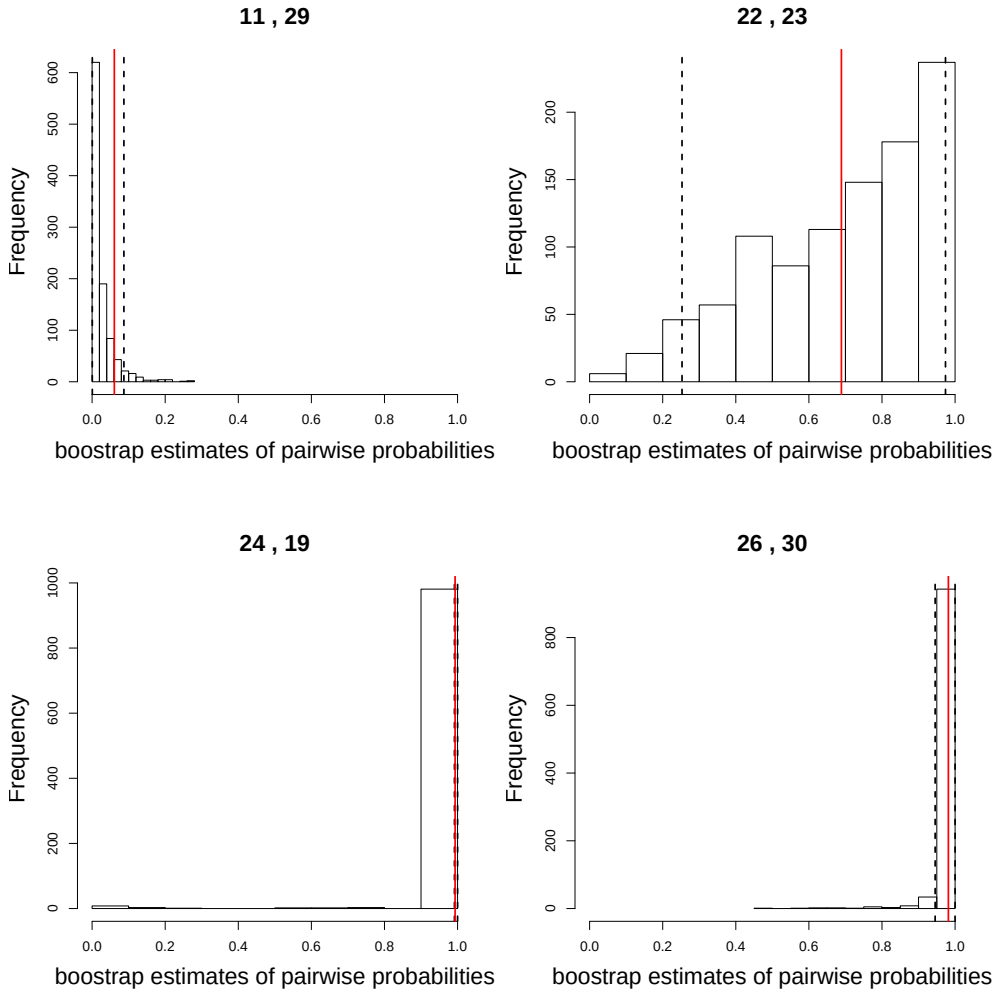


Figure 4: Histograms of bootstrap estimates $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$, $b = 1, \ldots, 1000$ for four pairs of objects (i, j) in the dataset of Example 3 (see Figure 3). The black broken vertical lines are the 0.025 and 0.975 quantiles P_{ij}^l and P_{ij}^u. The red solid vertical line is the true probability $P_{ij}(\boldsymbol{\theta})$. (This figure is better viewed in color).

Using the equalities $Bel_{ij}(\{s_{ij}\}) = m_{ij}(\{s_{ij}\})$ and $Pl_{ij}(\{s_{ij}\}) = 1 - Bel_{ij}(\{\neg s_{ij}\}) = 1 - m_{ij}(\{\neg s_{ij}\})$, we get

$$J(M) = \sum_{i<j} \left(m_{ij}(\{s_{ij}\}) - P_{ij}^l\right)^2 + \left(m_{ij}(\{\neg s_{ij}\}) - (1 - P_{ij}^u)\right)^2. \tag{9}$$

Assuming that $Bel_{ij}(\{s_{ij}\}) \approx P_{ij}^l$ and $Pl_{ij}(\{s_{ij}\}) \approx P_{ij}^u$, we will have, from (6),

$$\mathbb{P}\left(Bel_{ij}(\{s_{ij}\}) \leq P_{ij}(\boldsymbol{\theta}) \leq Pl_{ij}(\{s_{ij}\})\right) \approx 1 - \alpha. \tag{10}$$

Eq. (10) corresponds to the definition of a *predictive belief function* at confidence level $1 - \alpha$ as introduced in [11]. It is a particular kind of frequency-calibrated belief function as reviewed in [17].

To find an evidential partition M minimizing (9), let us assume that each mass function m_i has at most f nonempty focal sets in $\mathcal{F} = \{F_1, \ldots, F_f\} \subseteq 2^\Omega$. If c is small, we can take $\mathcal{F} = 2^\Omega \setminus \{\emptyset\}$. Otherwise, we can restrict the focal sets to have a cardinality less than some value (typically, 2). Each mass function m_i can then be represented by the f-vector $\boldsymbol{m}_i = (m_i(F_1), \ldots, m_i(F_f))^T$. Let $\boldsymbol{S} = (S_{k\ell})$ and $\boldsymbol{C} = (C_{k\ell})$ be the $f \times f$ matrices with general terms

$$S_{k\ell} := \begin{cases} 1 & \text{if } k = \ell \text{ and } |F_k| = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

and

$$C_{kl} := \begin{cases} 1 & \text{if } F_k \cap F_l = \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Furthermore, let $\boldsymbol{B}$ be the $2f \times f$ block matrix

$$\boldsymbol{B} := \begin{pmatrix} \boldsymbol{S} \\ \boldsymbol{C} \end{pmatrix},$$

and let $\boldsymbol{A}_j$ be the $2 \times 2f$ matrix defined by

$$\boldsymbol{A}_j := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \boldsymbol{m}_j^T, \tag{13}$$

where $\otimes$ is the Kronecker product.

With these notations, from (2), we have $m_{ij}(\{s_{ij}\}) = \boldsymbol{m}_j^T \boldsymbol{S} \boldsymbol{m}_i$, $m_{ij}(\{\neg s_{ij}\}) = \boldsymbol{m}_i^T \boldsymbol{C} \boldsymbol{m}_i$, and

$$\boldsymbol{m}_{ij} = \boldsymbol{A}_j \boldsymbol{B} \boldsymbol{m}_i, \tag{14}$$

with $\boldsymbol{m}_{ij} = (m_{ij}(\{s_{ij}\}), m_{ij}(\{\neg s_{ij}\}))^T$. Eq. (9) can thus be rewritten as

$$J(M) = \sum_{i<j} (\boldsymbol{m}_{ij} - \boldsymbol{m}_{ij}^*)^T (\boldsymbol{m}_{ij} - \boldsymbol{m}_{ij}^*) \tag{15a}$$

$$= \sum_{i<j} (\boldsymbol{A}_j \boldsymbol{B} \boldsymbol{m}_i - \boldsymbol{m}_{ij}^*)^T (\boldsymbol{A}_j \boldsymbol{B} \boldsymbol{m}_i - \boldsymbol{m}_{ij}^*), \tag{15b}$$

with $\boldsymbol{m}_{ij}^* = (P_{ij}^l, 1 - P_{ij}^u)^T$. From (15b), we can see that $J(M)$ is a quadratic function of $\boldsymbol{m}_i$. We can then use the Iterative Row-wise Quadratic Programming (IRQP) algorithm introduced by [47]. The IRQP is a block cyclic coordinate descent procedure [2, Section 2.7] minimizing $J(M)$ with respect to each vector $\boldsymbol{m}_i$ one at a time, while keeping the other vectors $\boldsymbol{m}_j$ for $j \neq i$ fixed. At each iteration, we thus minimize

$$J_i(\boldsymbol{m}_i) := \sum_{\substack{j=1 \\ j \neq i}}^{n} (\boldsymbol{A}_j \boldsymbol{B} \boldsymbol{m}_i - \boldsymbol{m}_{ij}^*)^T (\boldsymbol{A}_j \boldsymbol{B} \boldsymbol{m}_i - \boldsymbol{m}_{ij}^*) \tag{16a}$$

subject to

$$\mathbf{1}^T \boldsymbol{m}_i = 1 \quad \text{and} \quad \boldsymbol{m}_i \geq \mathbf{0}. \tag{16b}$$

Developing the expression in the right-hand side of (16a), we get

$$J_i(\boldsymbol{m}_i) = \boldsymbol{m}_i^T \boldsymbol{Q}_i \boldsymbol{m}_i + \boldsymbol{u}_i^T \boldsymbol{m}_i + a_i \tag{17}$$

with

$$\boldsymbol{Q}_i := \boldsymbol{B}^T \left(\sum_{j \neq i} \boldsymbol{A}_j^T \boldsymbol{A}_j \right) \boldsymbol{B} \tag{18a}$$

$$\boldsymbol{u}_i := -2 \left(\sum_{j \neq i} (\boldsymbol{m}_{ij}^*)^T \boldsymbol{A}_j \right) \boldsymbol{B} \tag{18b}$$

$$a_i := \sum_{j \neq i} (\boldsymbol{m}_{ij}^*)^T \boldsymbol{m}_{ij}^*. \tag{18c}$$

The minimization of function J_i in (17) subject to constraints (16b) can be performed using a standard quadratic programming solver. To define a stopping criterion, we compute a running mean of the relative error as follows: $e^{(0)} = 1$ and

$$e^{(t)} := \rho \, e^{(t-1)} + (1 - \rho) \frac{|J^{(t)} - J^{(t-1)}|}{J^{(t-1)}}, \quad t = 1, 2, \ldots, \tag{19}$$

where t is the iteration counter, $J^{(t)}$ is the value of the cost function at iteration t, and $\rho = 0.5$. The algorithm is then stopped when $e^{(t)} < \epsilon$, for some threshold ϵ. The whole procedure is summarized in Algorithm 2.

As matrix $\boldsymbol{Q}_i$ in (18a) is positive definitive, the quadratic programming problem (16) is convex [48] and has a unique solution. Consequently, the whole block coordinate descent procedure is guaranteed to converge to a local minimum [2, Proposition 2.7.1].

Example 4. *The procedure described in this section was applied to the data and bootstrap confidence intervals of Example 3. The set $\mathcal{F}$ of focal sets was defined to contain the singletons and the pairs, i.e.,*

$$\mathcal{F} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}\},$$

Algorithm 2 IRQP algorithm.

Require: Confidence intervals $\boldsymbol{m}_{ij}^{*}$ for $1 \leq i \leq j \leq n$, number of clusters c, focal sets $\mathcal{F} = \{F_1, \ldots, F_f\}$, stopping threshold ϵ

Initialize the evidential partition M randomly
$t := 0$, $e^{(0)} := 1$
Compute $J^{(0)}$ using (15)
while $e^{(t)} \geq \epsilon$ **do**
  $t := t + 1$
  $J^{(t)} := 0$
  for $i = 1$ **to** n **do**
    Compute $\boldsymbol{Q}_i$ and $\boldsymbol{u}_i$ in (18)
    Find $\boldsymbol{m}_i^{(t)}$ by minimizing (17) subject to (16b)
    Update M with $\boldsymbol{m}_i^{(t)}$
    $J^{(t)} := J^{(t)} + J_i(\boldsymbol{m}_i^{(t)})$
  end for
  $e^{(t)} := 0.5\, e^{(t-1)} + 0.5|J^{(t)} - J^{(t-1)}|/J^{(t-1)}$
end while

Ensure: Evidential partition M

Table 1: Mass functions for six objects displayed in Figure 6. For each object, the largest mass is printed in bold.

Object	$m(\{\omega_1\})$	$m(\{\omega_2\})$	$m(\{\omega_3\})$	$m(\{\omega_1, \omega_2\})$	$m(\{\omega_1, \omega_3\})$	$m(\{\omega_2, \omega_3\})$
3	0	0.042	0.113	0	0	**0.845**
4	0	0	**0.926**	0	0	0.074
10	0	0.406	0.007	0	0	**0.587**
11	**0.927**	0	0	0.073	0	0
12	0	**0.635**	0.005	0	0	0.360
22	0	0	0.141	0.092	**0.415**	0.352

and $f = 6$. Figure 5 shows the pairwise degrees of belief $Bel_{ij}(\{s_i\})$ and plausibility $Pl_{ij}(\{s_i\})$ as functions of, respectively, the lower bounds P_{ij}^{l} and the lower bounds P_{ij}^{u} of the bootstrap percentile 90% confidence intervals. Figure 6 presents a view of the resulting evidential partition, showing the maximum-plausibility hard partition as well as the convex hulls of the lower and upper approximations of each cluster [33]. These approximations are obtained by first assigning each object i to the set of clusters $A_i \subseteq \Omega$ with the highest mass, and then computing the lower and upper approximation defined by (1)*. The lower approximation of cluster k contains the objects that* surely *belong to that cluster, while the upper approximation contain those objects that* possibly *belong to cluster k. We can see that objects 3, 10 and 22 are ambiguous. Their mass functions, as well as those of three other objects are shown in Table 1.*

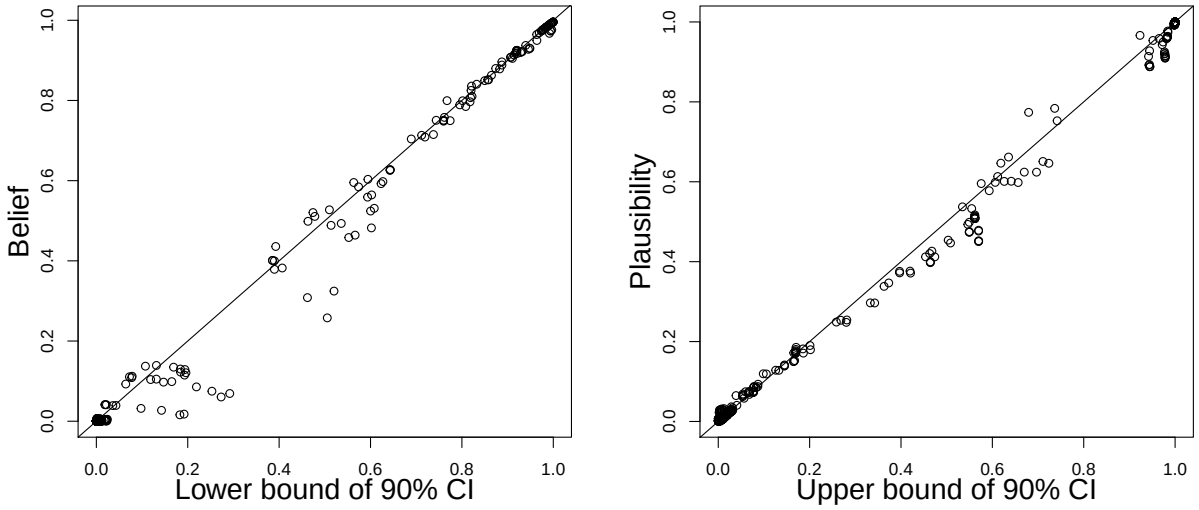


Figure 5: Pairwise degrees of belief $Bel_{ij}(\{s_i\})$ (left) and plausibility $Pl_{ij}(\{s_i\})$ (right) as functions of, respectively, the lower bounds P_{ij}^l and the lower bounds P_{ij}^u of bootstrap percentile 90% confidence intervals, for the data of Example 4.

3.4. Complexity analysis

The propose clustering methods consists in three steps:

1. The computation of the estimates $\widehat{\boldsymbol{\theta}}_b$ and $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$ for each of the B bootstrap samples (lines 1-7 in Algorithm 1)
2. The computation of the quantiles P_{ij}^l and P_{ij}^u (lines 8-11 in Algorithm 1);
3. The construction of the evidential partition (Algorithm 2).

In Step 1, each iteration of the EM has complexity $O(cn)$. Assuming that the number of iterations is roughly constant and does not depend on n, the computation of each estimate $\boldsymbol{\theta}_b$ has complexity $O(cn)$, and the computation of $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$ for all $i < j$ involves $O(cn^2)$ operations. So, the complexity of Step 1 is $O(Bcn^2)$. In Step 2, each quantile can be computed in $O(n)$ operations [5], so the complexity of Step 2 is $O(Bn^2)$. Finally, solving each quadratic programming problem in Step 3 has worst-case complexity $O(f^3)$, where f is the number of focal sets, so that each iteration of Algorithm 2 has $O(nf^3)$ complexity. Assuming the number of iterations of the IRQP algorithm to be roughly constant, the complexity of Step 3 is $O(nf^3)$. Overall, the time complexity of the global procedure is $O(Bcn^2 + nf^3)$. As far as storage space is concerned, we need to store the confidence intervals, which has $O(n^2)$ space complexity, and the evidential partition, which takes $O(nf)$ space, so that the overall complexity is $O(n^2 + nf)$.

In the worst case, the number of nonempty focal sets is $2^c - 1$. It is thus crucial to limit the number of focals sets when c is large. A simple strategy is to restrict the focal sets

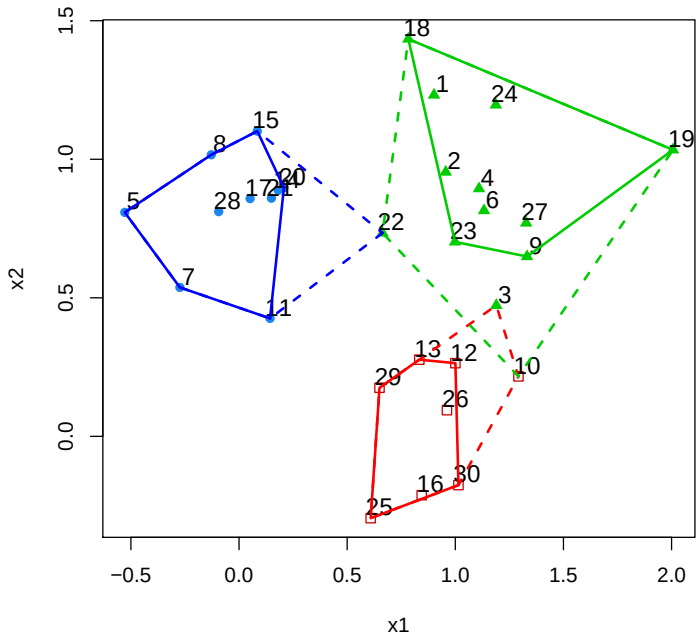


Figure 6: Evidential partition of the data of Example 4. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster, as defined in the text. The lower approximation of a cluster contains the objects that can confidently be assigned to that cluster, while the upper approximation contains objects that may belong to several clusters. For instance, Object 22 may belong to clusters ω_1 (left) or ω_3 (top right), while object 3 may clusters ω_2 (bottom right) or ω_3 (bottom right).

of mass functions m_i in the evidential partition to singletons and pairs, which bring their number down to $c(c+1)/2$. A more sophisticated strategy, proposed in [14] is to first identify the pairs of overlapping clusters, and to use only these pairs (as well as the singletons) as focal sets; this strategy will be illustrated in Section 4.2 with the GvHD dataset.

Another limitation of our method is its $O(n^2)$ complexity, which precludes application to very large datasets. We can remark that our approach is especially useful with small and medium-size datasets (typically, containing a few hundred or thousand objects), for which the cluster-membership probabilities usually cannot be estimated accurately. Nevertheless, some preliminary ideas to make our approach applicable to large datasets will be mentioned in the last paragraph of Section 5 as directions for future work.

4. Experimental results

We first present results with simulated data in Section 4.1 to verify the calibration property experimentally. Some results with real datasets are then reported in Section 4.2. All the simulations reported in this section were performed using an implementation of our algorithm in R publicly available as function `bootclus` in package `evclust` [16].

4.1. Simulated data

We first considered datasets with $n = 300$ observations drawn from three different two-dimensional Gaussian mixture models (GMM) with $c = 3$ components and the following parameters:

Mixture 1:

$$\boldsymbol{\mu}_1 := (0,0)^T, \quad \boldsymbol{\mu}_2 := (0,3)^T, \quad \boldsymbol{\mu}_3 := (3,0)^T,$$
$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \pi_1 = \pi_2 = \pi_3 := 1/3.$$

Mixture 2:

$$\boldsymbol{\mu}_1 := (0,0)^T, \quad \boldsymbol{\mu}_2 := (0,2.5)^T, \quad \boldsymbol{\mu}_3 := (2.5,0)^T,$$
$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 := \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \pi_1 = \pi_2 = \pi_3 := 1/3.$$

Mixture 3:

$$\boldsymbol{\mu}_1 := (0,0)^T, \quad \boldsymbol{\mu}_2 := (0,3)^T, \quad \boldsymbol{\mu}_3 := (3,0)^T,$$
$$\boldsymbol{\Sigma}_1 := \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 := 1.5 \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_3 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$
$$\pi_1 = \pi_2 = \pi_3 := 1/3.$$

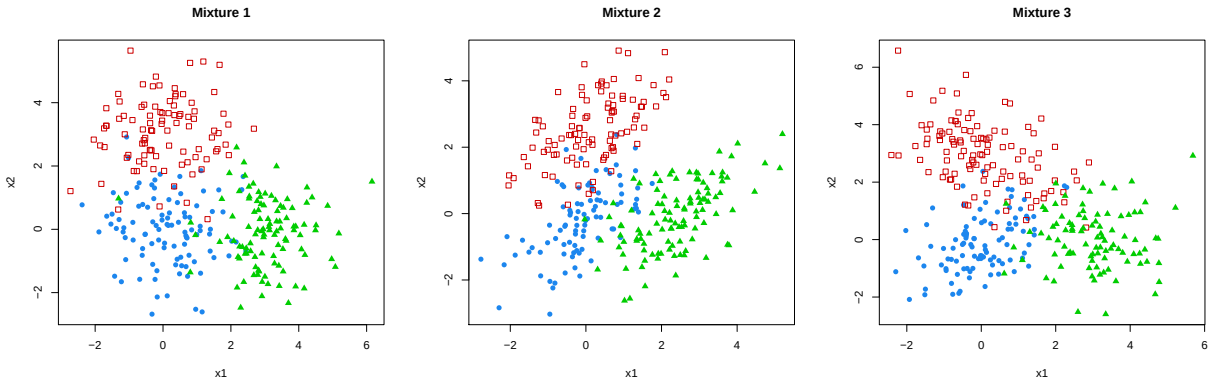


Figure 7: Three datasets drawn from three Gaussian mixtures with $c = 3$ components.

We generated 100 datasets from each distribution. Examples of datasets are shown in Figure 7.

For each dataset, we generated $B = 1000$ nonparametric bootstrap samples and we estimated the parameters of three-component GMMs under four assumptions[1]:

1. Spherical distributions, equal volume (EII);
2. Ellipsoidal distributions, equal volume, shape, and orientation (EEE);
3. Ellipsoidal distributions, varying volume, shape, and orientation (VVV);
4. Best model according to the BIC criterion (Auto).

Here, the terms "volume", "shape" and "orientation" refer to the eigenvalue decomposition of covariance matrices:

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T,$$

where parameter $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/d}$, $\boldsymbol{D}_k$ is a matrix with eigenvectors, and $\boldsymbol{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues of $\boldsymbol{\Sigma}_k$, scaled such that $|\boldsymbol{A}_k| = 1$. With this parameterization, each of the three sets of parameters has a geometrical interpretation: λ_k indicates the volume of cluster k, $\boldsymbol{D}_k$ its orientation, and $\boldsymbol{A}_k$ its shape [1].

It is clear that EII, EEE and VVV are the exact models for, respectively, Mixtures 1, 2 and 3. When the model selection strategy was employed, we selected the best model on the whole dataset, and we fitted the selected model on each bootstrap replicate. For each dataset and each model, we computed bootstrap confidence intervals $[P_{ij}^l, P_{ij}^u]$ on $P_{ij}(\boldsymbol{\theta})$ for each pair of objects (i, j) using Algorithm 1, at confidence levels $\alpha = 0.1$ and $\alpha = 0.05$.

Examples of 90% confidence intervals and approximating belief and plausibility degrees for four object pairs in one particular dataset drawn from Mixture 2 are shown in Figure 8. In these four examples, both intervals $[P_{ij}^l, P_{ij}^u]$ and $[Bel_{ij}(\{s_{ij}\}), Pl_{ij}(\{s_{ij}\})]$ contain the true probability $P_{ij}(\boldsymbol{\theta})$ that object i and j are in the same class. Figure 9 plots the belief and plausibility degrees $Bel_{ij}(\{s_{ij}\})$ and $Pl_{ij}(\{s_{ij}\})$ vs. the lower and upper bounds

[1]We used the R package `mclust` [43].

of 90% confidence intervals on $P_{ij}(\boldsymbol{\theta})$. We can see that there is a reasonably good fit between the belief-plausibility intervals and the bootstrap confidence intervals, thanks to the minimization of criterion $J(M)$ in (9). The belief and plausibility degrees are plotted against the true probabilities $P_{ij}(\boldsymbol{\theta})$ in Figure 10. For this dataset, almost all the belief-plausibility intervals contained the true probabilities.

Tables 2-4 show the estimated coverage probabilities (i.e., the proportion of intervals containing the true value $P_{ij}(\boldsymbol{\theta})$ for 90% and 95% confidence intervals and their approximations by belief-plausibility intervals. We can see that confidence intervals and belief-plausibility intervals have similar coverage probabilities, and these probabilities are close to their nominal levels *when the model is correctly specified*. For instance, in Table 2, the true model is EII, which is a special case of models EEE and VVV. Consequently, all three models are correct in this case, and they lead to intervals with coverage probability close to the specified value. However, assuming a more general model such as VVV results in wider intervals because of the larger standard error of the estimates. When the true model is EEE (Table 3), assuming the incorrect model EII has a devastating effect in terms of coverage probabilities, which are then much smaller than the specified level. The same phenomenon is observed in Table 4, where the correct model is VVV and models EII and EEE are both wrong. The automatic model determination method works well when the true model is EII or EEE (Tables 2 and 3), but it does not work so well when the true model is VVV (Table 4), because it sometimes select a simpler model than the true one.

From these experiments, we can conclude that the belief-plausibility intervals have coverage probabilities close to their nominal confidence levels when a correct model is assumed. Correct assumptions about parameter constraints (such as homoscedasticity) make it possible to obtain shorter intervals when the assumptions are correct, but their can have a negative effect on coverage probabilities when the assumptions are wrong. Automatic model selection based, e.g., on the BIC criterion can be used, but the selection should be biased in favor of more complex models to avoid model misspecification.

Experiment with non-normal data. Given the importance of correct model specification to ensure the frequency-calibration of belief-plausibility intervals, we can expect poor results when fitting a GMM to data generated by a mixture whose components are significantly non-normal. As a case study, we considered data from a mixture of three two-dimensional skew t distributions [49] with the following parameters:

$$\boldsymbol{\mu}_1 := (3,-4)^T, \quad \boldsymbol{\mu}_2 := (3.5,4)^T, \quad \boldsymbol{\mu}_3 := (2,2)^T,$$

$$\boldsymbol{\Sigma}_1 := \begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\pi_1 = \pi_2 := 0.4, \quad \pi_3 := 0.4.$$

$$\nu_1 := 3, \quad \nu_2 = \nu_3 := 5$$

$$\boldsymbol{\delta}_1 := (3,3)^T, \quad \boldsymbol{\delta}_2 := (1,5)^T, \quad \boldsymbol{\delta}_3 := (-3,1)^T,$$

where ν_k and $\boldsymbol{\delta}_k$ denote, respectively, the degrees of freedom and the skewness parameters.

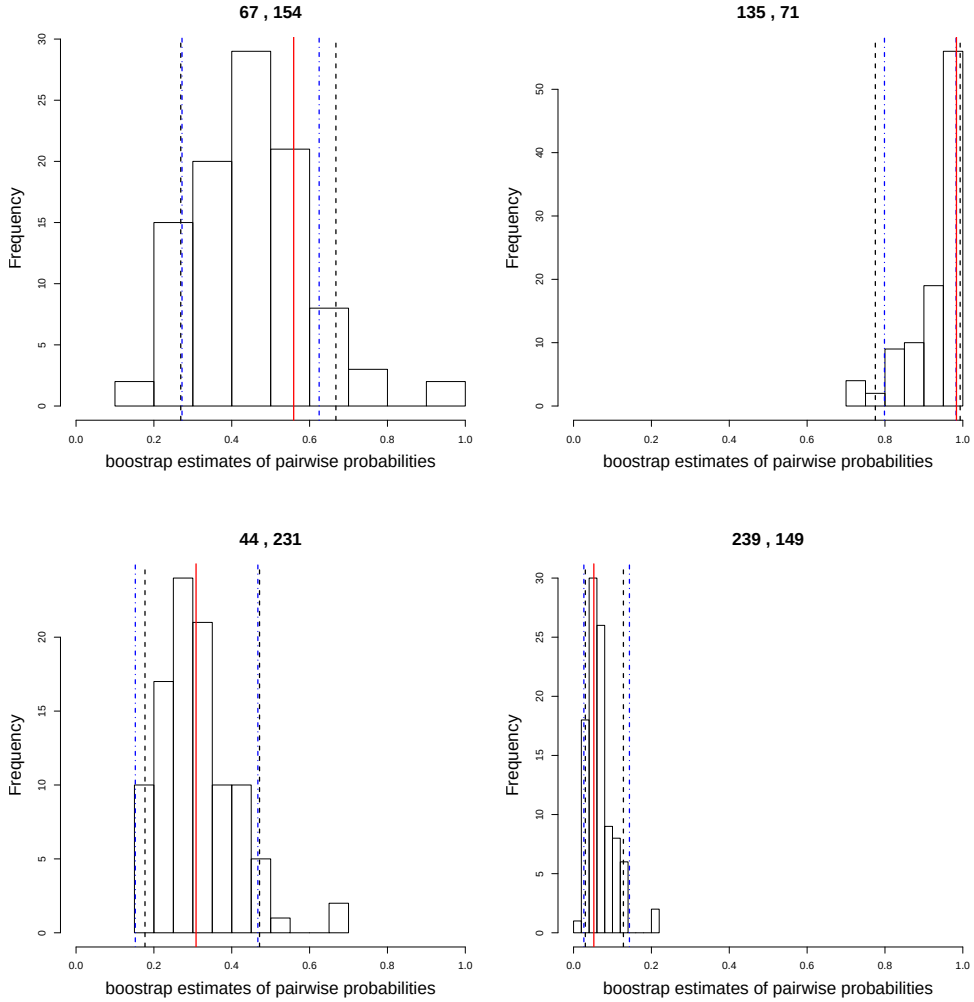
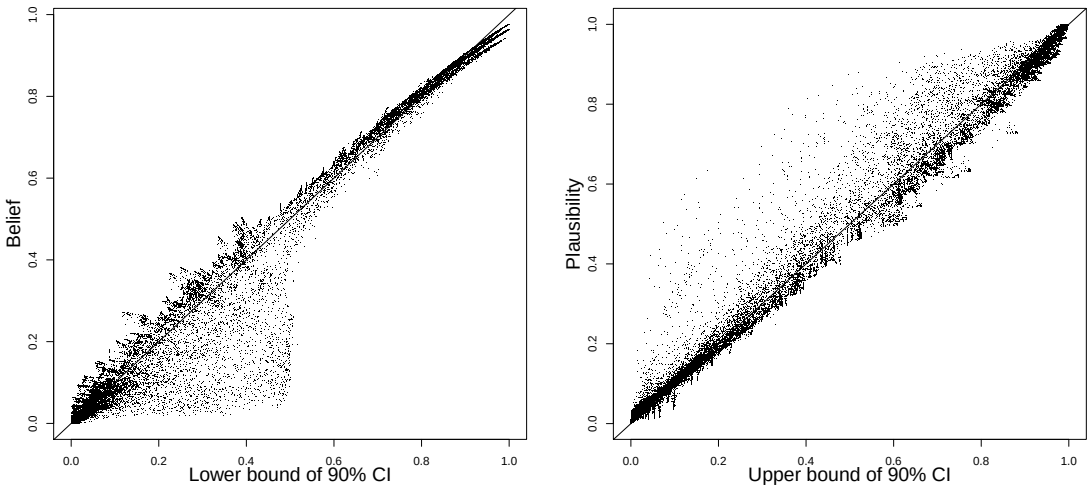
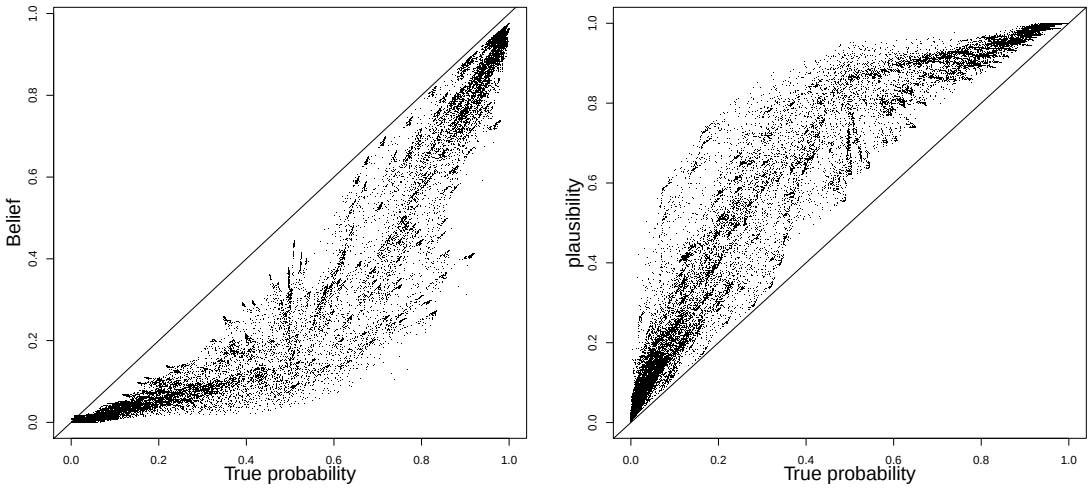


Figure 8: Histograms of bootstrap estimates $P_{ij}(\widehat{\boldsymbol{\theta}}_b)$, $b = 1, \ldots, 1000$ for four pairs of objects (i, j) in a particular dataset drawn from Mixture 2. The black broken vertical lines are the 0.05 and 0.95 quantiles P_{ij}^l and P_{ij}^u. The blue dash-dot vertical lines are belief and plausibility degrees $Bel_{ij}(\{s_{ij}\})$ and $Pl_{ij}(\{s_{ij}\})$. The red solid vertical line is the true probability $P_{ij}(\boldsymbol{\theta})$. (This figure is better viewed in color).



(a) (b)

Figure 9: Dataset drawn from Mixture 2: (a) Lower bound P_{ij}^l of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) Upper bound P_{ij}^u of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).



(a) (b)

Figure 10: Dataset drawn from Mixture 2: (a) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).

Table 2: Coverage rates and lengths of bootstrap confidence intervals (CI) and belief-plausibility intervals for 100 datasets generated from Mixture 1 (model EII), at nominal 90% and 95% confidence levels. The numbers in parentheses are the standard deviations over the 100 datasets. The coverage rates for correctly specified models are printed in bold.

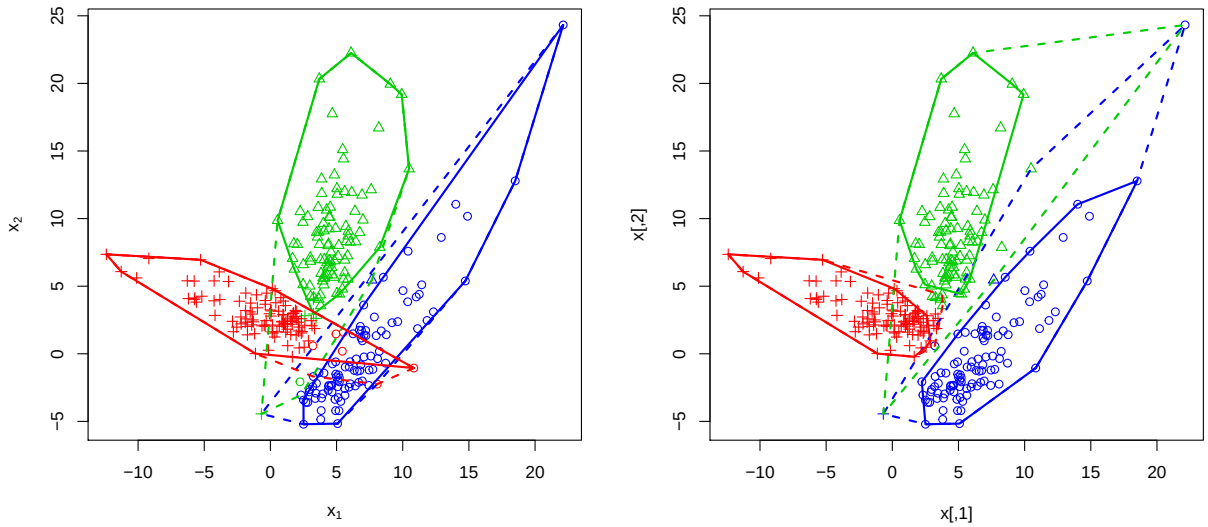
			Assumed model							
			EII		EEE		VVV		Auto	
True model	$1-\alpha$		CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]
EII	0.90	cover.	**0.87**	**0.90**	**0.88**	**0.90**	**0.92**	**0.91**	0.88	0.89
			(0.159)	(0.101)	(0.125)	(0.091)	(0.102)	(0.088)	(0.15)	(0.12)
		length	0.11	0.11	0.14	0.14	0.32	0.32	0.11	0.11
			(0.017)	(0.017)	(0.028)	(0.028)	(0.085)	(0.088)	(0.018)	(0.018)
	0.95	cov.	**0.93**	**0.94**	**0.94**	**0.94**	**0.96**	**0.94**	0.93	0.93
			(0.121)	(0.079)	(0.087)	(0.065)	(0.067)	(0.062)	(0.126)	(0.097)
		length	0.13	0.13	0.17	0.17	0.39	0.40	0.133	0.132
			(0.021)	(0.021)	(0.033)	(0.033)	(0.097)	(0.100)	(0.022)	(0.022)

Table 3: Coverage rates and lengths of bootstrap confidence intervals (CI) and belief-plausibility intervals for 100 datasets generated from Mixture 2 (model EEE), at nominal 90% and 95% confidence levels. The numbers in parentheses are the standard deviations over the 100 datasets. The coverage rates for correctly specified models are printed in bold.

			Assumed model							
			EII		EEE		VVV		Auto	
True model	$1-\alpha$		CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]
EEE	0.90	cover.	0.34	0.50	**0.89**	**0.91**	**0.89**	**0.89**	0.88	0.90
			(0.033)	(0.038)	(0.122)	(0.080)	(0.125)	(0.114)	(0.155)	(0.107)
		length	0.16	0.16	0.15	0.15	0.37	0.37	0.16	0.16
			(0.032)	(0.032)	(0.031)	(0.031)	(0.082)	(0.084)	(0.036)	(0.036)
	0.95	cov.	0.40	0.56	**0.95**	**0.95**	**0.95**	**0.92**	0.94	0.94
			(0.035)	(0.040)	(0.085)	(0.056)	(0.088)	(0.086)	(0.112)	(0.083)
		length	0.19	0.19	0.18	0.19	0.45	0.46	0.19	0.19
			(0.038)	(0.039)	(0.037)	(0.037)	(0.088)	(0.091)	(0.043)	(0.044)

Table 4: Coverage rates and lengths of bootstrap confidence intervals (CI) and belief-plausibility intervals for 100 datasets generated from Mixture 3 (model VVV), at nominal 90% and 95% confidence levels. The numbers in parentheses are the standard deviations over the 100 datasets. The coverage rates for correctly specified models are printed in bold.

			Assumed model							
			EII		EEE		VVV		Auto	
True model	$1-\alpha$		CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]	CI	[Bel,Pl]
VVV	0.90	cover.	0.47	0.58	0.57	0.64	**0.90**	**0.89**	0.65	0.70
			(0.078)	(0.077)	(0.136)	(0.139)	(0.126)	(0.110)	(0.195)	(0.162)
		length	0.16	0.16	0.24	0.25	0.31	0.32	0.18	0.18
			(0.039)	(0.040)	(0.083)	(0.087)	(0.080)	(0.083)	(0.037)	(0.037)
	0.95	cov.	0.55	0.65	0.67	0.73	**0.95**	**0.93**	0.74	0.78
			(0.080)	(0.079)	(0.128)	(0.124)	(0.089)	(0.077)	(0.177)	(0.147)
		length	0.19	0.20	0.30	0.31	0.39	0.40	0.22	0.22
			(0.046)	(0.047)	(0.093)	(0.097)	(0.097)	(0.099)	(0.046)	(0.047)

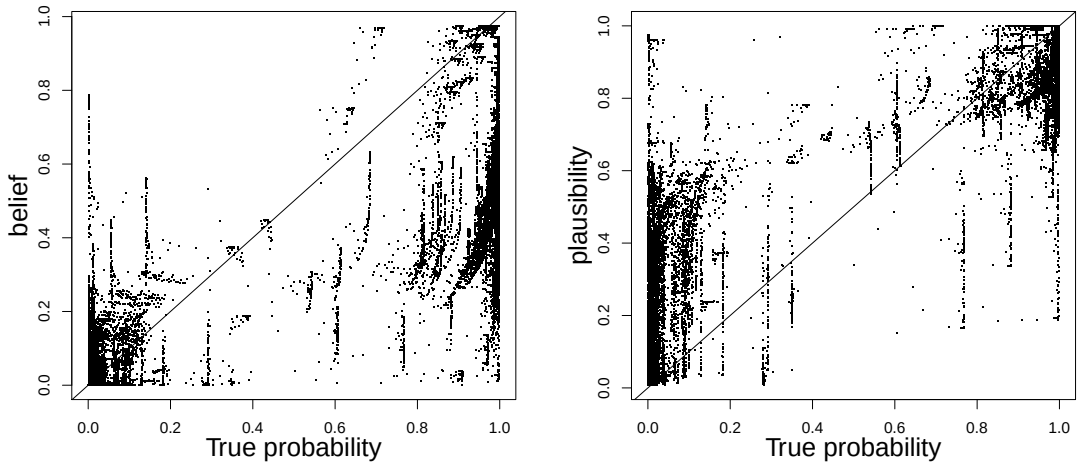


(a) (b)

Figure 11: Evidential partitions of a dataset drawn from a mixture of skew t distributions, fitted with a GMM (a) and with a mixture of skew t distributions (b). The true groups are represented by different symbols, and the maximum-plausibility groups are represented by different colors. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

Figure 11a shows a dataset of $n = 300$ observations drawn from this distribution, together with the partition obtained by fitting a GMM with the assumption of equal volume of the three clusters (model EVV in package `mclust`), as well as the lower and upper approximations of each cluster. We can see that the partition obtained with the normality assumption is close to the true partition (with only 12 misclassified points out of 300). However, only 50.4% of the belief-plausibility intervals computed from 90% bootstrap confidence intervals contain the true probabilities, which suggests that their true coverage probability is significantly smaller than the nominal one (see Figure 12).

As noted by McLachlan and Basford [35, Section 2.7], "In the situation where the sample is completely unclassified, as in the usual cluster analysis setting where there is no genuine group structure, it is a difficult task to assess the fit of a mixture model". For assessing the fit of a GMM, a method that is not fully rigorous but that works reasonable well in practice is to fit a GMM first, and then to test the normality of the data in each cluster. Here, normality is rejected for all three components with high significance by, for instance, Henze-Zirkler's test of multivariate normality [25], with p-values equal to 3.2×10^{-5}, 4.0×10^{-7} and 3.9×10^{-5}. Figure 11b displays the obtained partition as well as the lower and upper approximations of each cluster obtained by fitting a mixture of skew t distributions to the data (using the R package `EMMIXskew` [50]). As shown by Figure 13, 93.2% of the belief-



(a) (b)

Figure 12: Dataset drawn from a mixture of skew t distributions fitted with a GMM: (a) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).

plausibility intervals now contain the true probabilities, which is close to the nominal value of 90%.

These results suggest that the belief-plausibility intervals computed by our method may not be well calibrated when there is a severe lack of fit of the mixture model to the data, even though the obtained credal partition can still reveal the clustering structure of the data. In most cases, however, the data distribution can be reasonably well approximated by a GMM. This model will be assumed for the analysis of real datasets carried out in the next section.

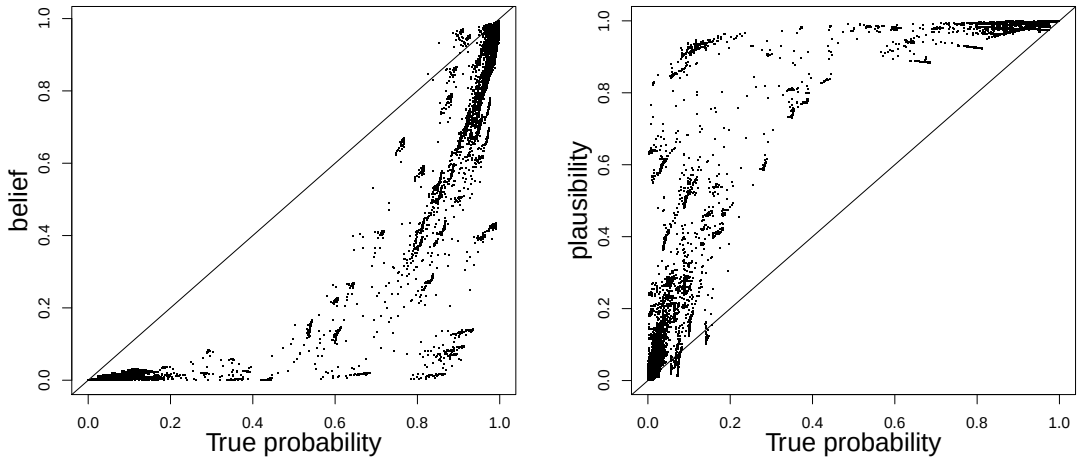
4.2. Real data

In this section, we apply our approach with GMMs to three real datasets, and we compare it to two evidential clustering algorithms: ECM [33] and EVCLUS [13, 14], both implemented in the R package `evclust` [16].

Iris data

We first consider the well-known Iris dataset[2], composed of 150 four-dimensional vectors partitioned in three groups corresponding to three species of Iris flowers (setosa, versicolor and virginica, abbreviated as se, ve and vi). For this dataset we fixed the number of clusters to $c = 3$, and we searched for the best GMM model using function `Mclust` in the `mclust` package. The selected model was "VEV" corresponding to ellipsoidal clusters with equal

[2]Available in the R package `datasets`.



(a) (b)

Figure 13: Dataset drawn from a mixture of skew t distributions, fitted with the correct model: (a) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) True probability $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).

shape. The result is represented graphically in Figure 14, showing the obtained partition as well as the cluster centers and cluster shapes represented by isodensity ellipses. The adjusted rand index (ARI) for the obtained partition is 0.90, with five objects from the *versicolor* group incorrectly assigned to the *virginica* group.

We then computed 90% bootstrap percentile confidence intervals using Algorithm 1 with $B = 1000$, and we constructed an evidential partition using Algorithm 2, with $f = 6$ focal sets (the singletons and the pairs). As shown by Figure 15, the confidence bounds are quite well approximated by the belief-plausibility intervals. Some belief values are smaller than the lower bounds of the confidence intervals (Figure 15b), which suggests that the coverage probability of these intervals might be larger than the 90% specified level.

The lower and upper cluster approximations for the obtained evidential partition are represented in Figure 16. We can see that the *setosa* group, which is well separated from the other two, has a precise representation (for that cluster, the lower and upper approximations are equal). In contrast, the other two groups are overlapping, resulting in some objects being assigned to more than one group. Table 5 shows the mass functions for the five objects from the *versicolor* group wrongly clustered with the *virginica* in the model-based clustering. We can see that four of them (objects 69, 71, 73 and 78) have a large mass on the set $\{\mathsf{ve}, \mathsf{vi}\}$ corresponding to the union of the *versicolor* and *virginica*, which indicates doubt in the assignment to any of these two clusters. Table 6 shows the confusion matrix, after each object has been assigned to the cluster subset with the highest mass. Clusters were labeled according to the majority group of objects they contained. We can see that 11 objects from the *versicolor* group, and three from the *virginica*, are assigned to the set $\{\mathsf{ve}, \mathsf{vi}\}$. Only

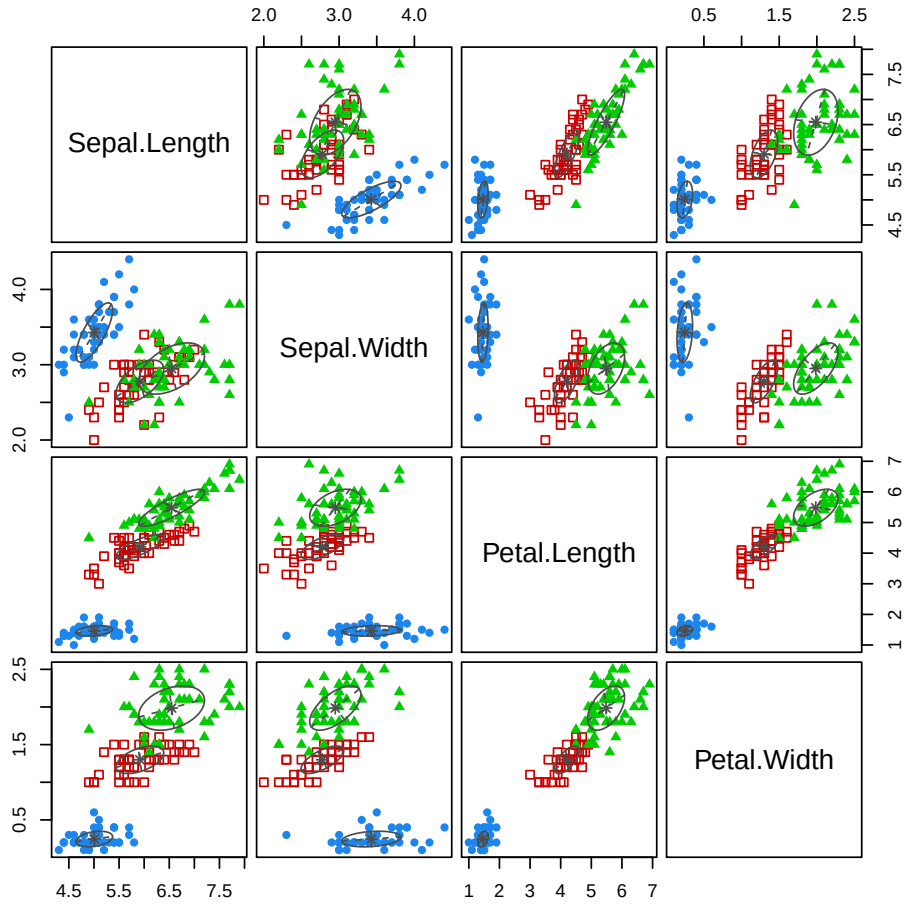
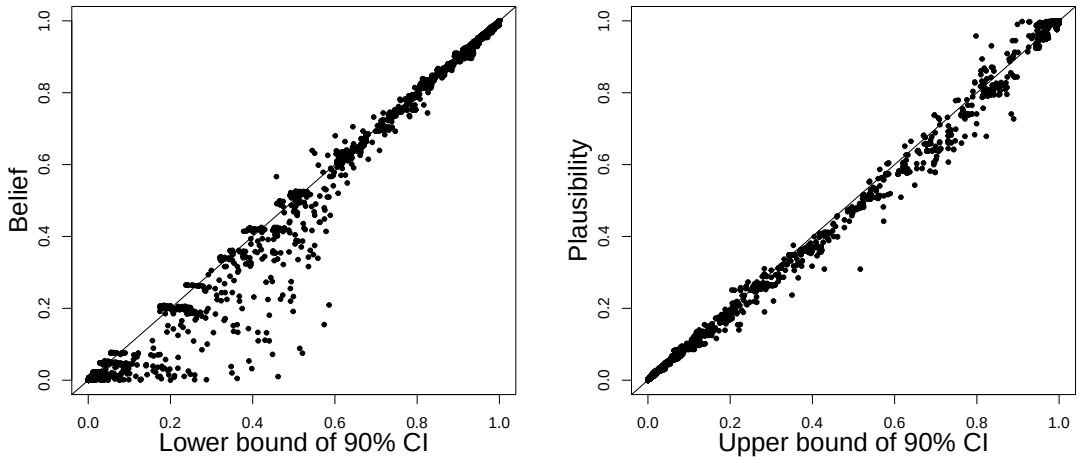


Figure 14: Iris data with the partition obtained by fitting a GMM with $c = 3$ components. Covariances in each group are represented by isodensity ellipses. (This figure is better viewed in color).



(a) (b)

Figure 15: Approximation of confidence intervals by belief-plausibility intervals for the Iris data. (a) Lower bound P_{ij}^l of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) Upper bound P_{ij}^u of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).

Table 5: Mass functions for the five misclassified instances in the Iris dataset. The three clusters have been renamed as se, ve and vi.

Object	$m(\{\mathsf{se}\})$	$m(\{\mathsf{ve}\})$	$m(\{\mathsf{vi}\})$	$m(\{\mathsf{se},\mathsf{ve}\})$	$m(\{\mathsf{se},\mathsf{vi}\})$	$m(\{\mathsf{ve},\mathsf{vi}\})$
69	0.012	0	0.007	0	0	0.991
71	0	0.005	0.077	0	0	0.918
73	0	0.003	0.202	0	0	0.795
78	0	0.051	0.052	0	0	0.897
84	0	0	0.882	0	0	0.117

objet (# 84) from the *versicolor* group is misclassified as *virginica*.

We also compared the above results to those obtained using ECM and EVCLUS. For ECM, we set the parameters α and β to their default values ($\alpha = 1$ and $\beta = 2$), and we set $\delta = 100$ to avoid having any mass on the empty set. To select the focal sets, we used the method described in [14]: we first ran the algorithm using only the singletons as focal sets, and we found the pairs of classes with high similarity (see [14] for details). Here, the pair $\{\mathsf{ve},\mathsf{vi}\}$ was selected. Then, we ran the ECM algorithm again with focal sets $\{\mathsf{se}\}$, $\{\mathsf{ve}\}$, $\{\mathsf{vi}\}$ and $\{\mathsf{ve},\mathsf{vi}\}$. The resulting evidential partition is shown in Figure 17, and the confusion matrix is shown in Table 7. As we can see, ECM tends to extract spherical clusters, and thus fails to identify correctly the *versicolor* and *virginica* groups. Comparing Tables 6 and 7, we can see that ECM also misclassified one *virginica* object as *versicolor*, but it provides a much more imprecise evidential partition, with 16 objects from the *versicolor* group and

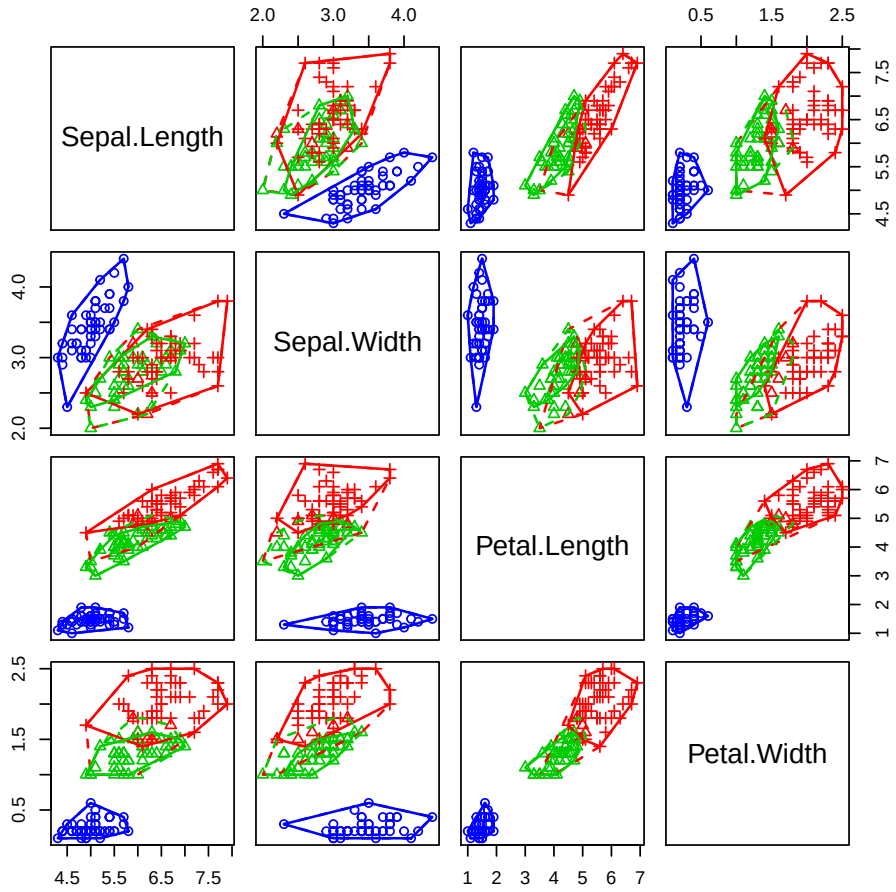


Figure 16: Evidential partition of the Iris data using the model-based approach. The true groups are represented by different symbols (o: setosa; triangle: versicolor; +: virginica), and the maximum-plausibility groups are represented by different colors. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

Table 6: Confusion matrix for the Iris dataset.

	Clustering			
	{se}	{ve}	{vi}	{ve, vi}
setosa	50	0	0	0
versicolor	0	38	1	11
virginica	0	0	47	3

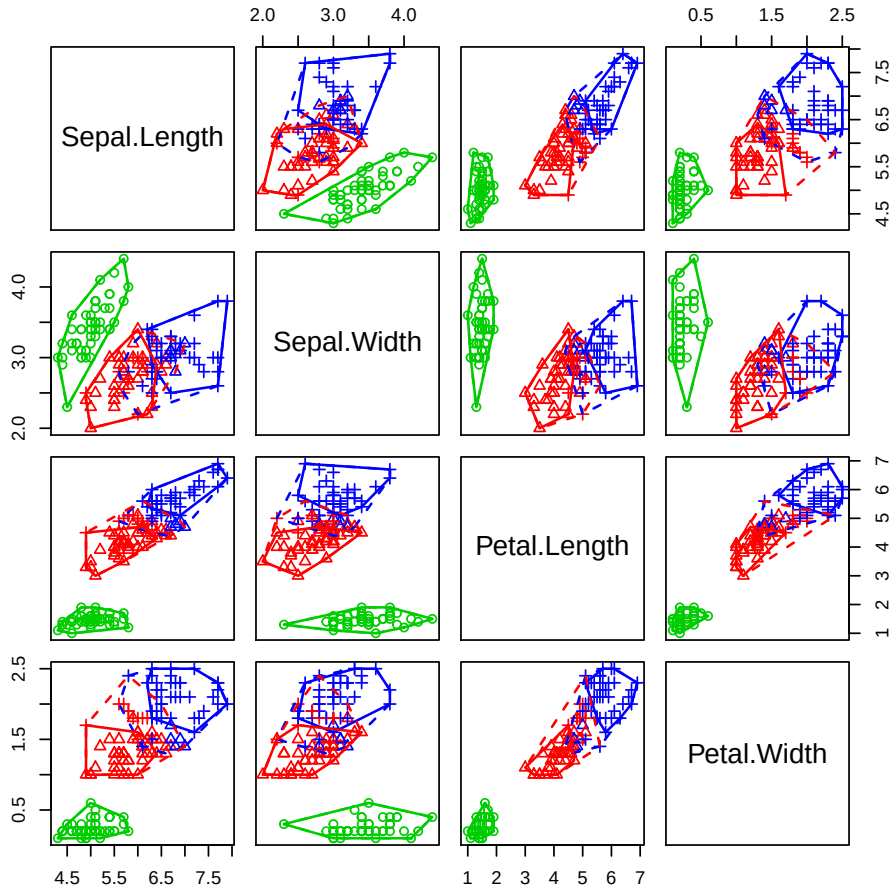


Figure 17: Evidential partition obtained by ECM applied to the Iris data. The true groups are represented by different symbols (o: setosa; triangle: versicolor; +: virginica), and the maximum-plausibility groups are represented by different colors. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

17 objects from the *virginica* assigned to the compound cluster $\{\mathsf{ve}, \mathsf{vi}\}$.

Finally, we also applied the k-EVCLUS algorithm [14] to the same data, after normalizing the four attributes. We used the default settings and $k = 50$ (see [14] for details). We used the same procedure as with ECM to identify pairs of clusters to include as focal sets, and the whole set $\Omega = \{\mathsf{se}, \mathsf{ve}, \mathsf{vi}\}$ was also included as a focal set. Again, the pair $\{\mathsf{ve}, \mathsf{vi}\}$ was correctly identified and included as focal set. The resulting evidential partition is shown in Figure 18, and the confusion matrix is shown in Table 8. EVCLUS is designed to assign some mass to the empty set, with a high mass on the empty set signaling an outlier. Here, four points were identified as outliers: they are the points outside the cluster upper approximations in Figure 8. As shown by the confusion matrix in Table 8, EVCLUS does not perform very well on this dataset, with roughly the same number of correctly classified

Table 7: Confusion matrix for the evidential partition obtained by ECM on the Iris dataset.

	Clustering			
	{se}	{ve}	{vi}	{ve, vi}
setosa	50	0	0	0
versicolor	0	34	0	16
virginica	0	1	32	17

Table 8: Confusion matrix for the evidential partition obtained by k-EVCLUS on the Iris dataset.

	Clustering				
	∅	{se}	{ve}	{vi}	{ve, vi}
setosa	2	48	0	0	0
versicolor	0	0	32	10	8
virginica	2	0	6	33	9

objects as ECM, but 16 misclassified objects. Overall, both ECM and EVCLUS performed significantly worse on this dataset than the model-based approach.

Diabetes data

The Diabetes dataset[3] [42, 43] contains three measurements made on 145 non-obese adult patients classified into three groups (normal, overt, and chemical, abbreviated as no, ov and ch). The three attributes are glucose (area under plasma glucose curve after a three hour oral glucose tolerance test), insulin (area under plasma insulin curve after a three hour oral glucose tolerance test), and sspg (steady state plasma glucose). For this dataset, the best model according to BIC was found to be the full unconstrained model ("VVV") with $c = 3$ components. The data with the obtained partition as well as the estimated cluster centers and covariance ellipses are shown in Figure 19. The ARI for the obtained partition is 0.66, and the confusion matrix is shown in Table 9. As before, clusters were labeled from the majority group among their elements. As we can see, there are 20 misclassified objects.

As before, we computed 90% bootstrap percentile confidence intervals with $B = 1000$, and we used these intervals to constructed an evidential partition with $f = 6$ focal sets (the singletons and the pairs). The resulting evidential partition is displayed in Figure 21

[3]Available in the R package `mclust`.

Table 9: Confusion matrix for the hard partition of the Diabetes dataset (model-based approach).

	Clustering		
	{ch}	{no}	{ov}
chemical	26	9	1
normal	4	72	0
overt	6	0	27

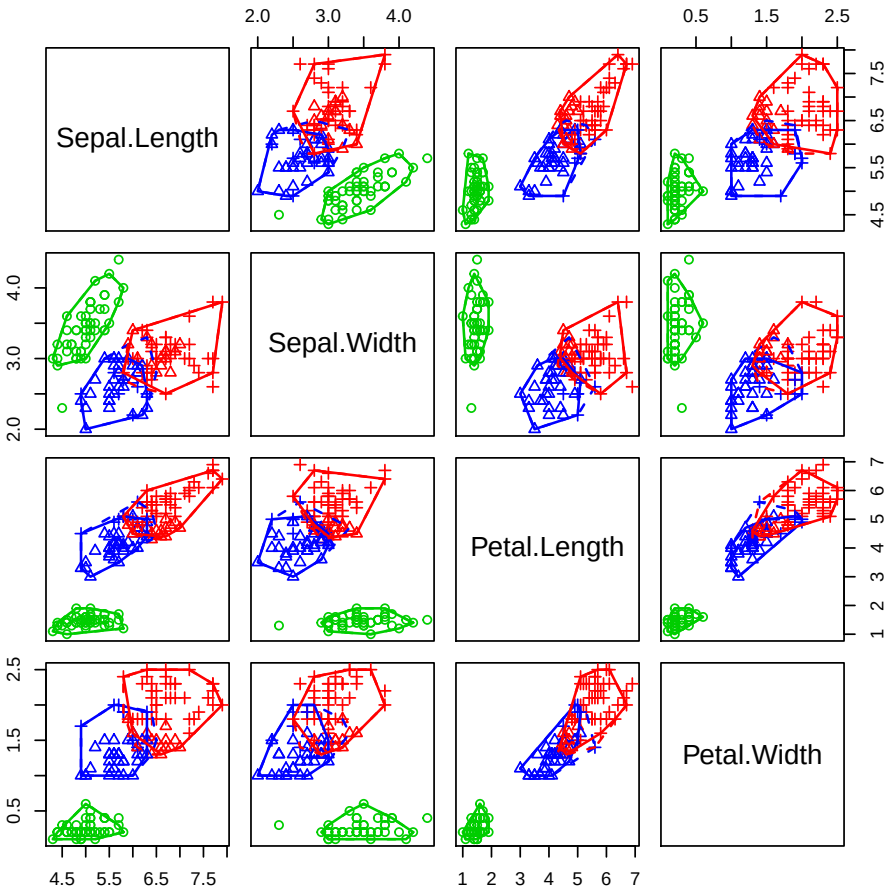


Figure 18: Evidential partition obtained by k-EVCLUS applied to the Iris data. The true groups are represented by different symbols (o: setosa; triangle: versicolor; +: virginica), and the maximum-plausibility groups are represented by different colors. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

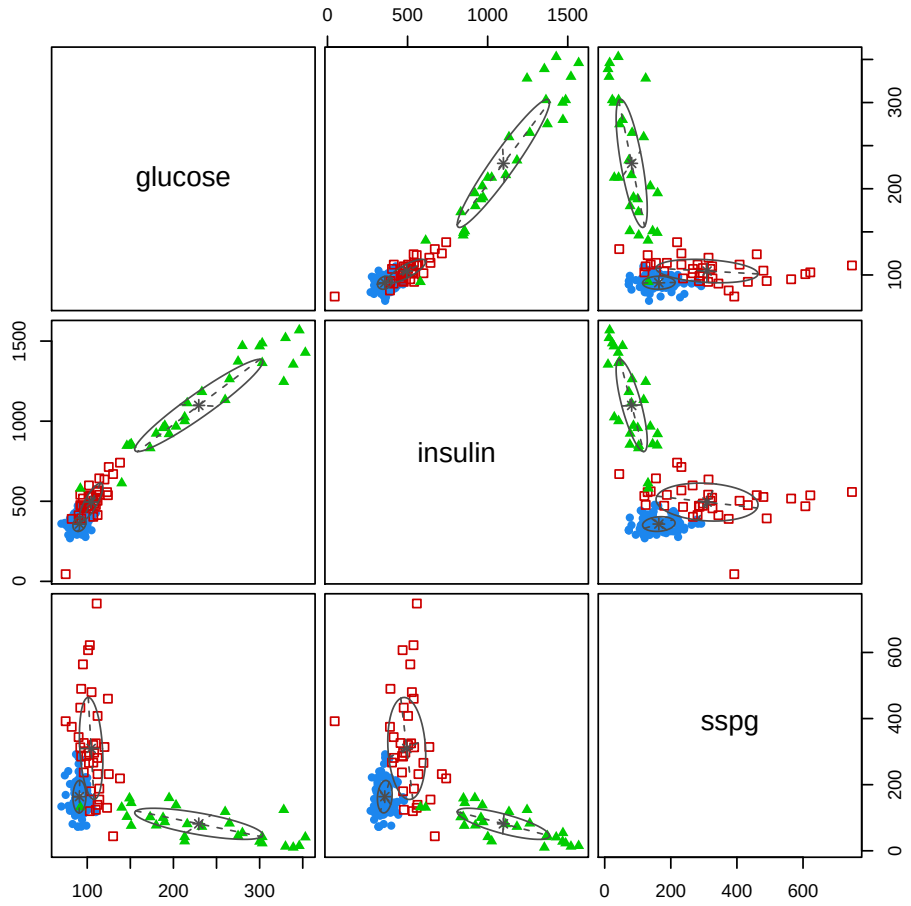
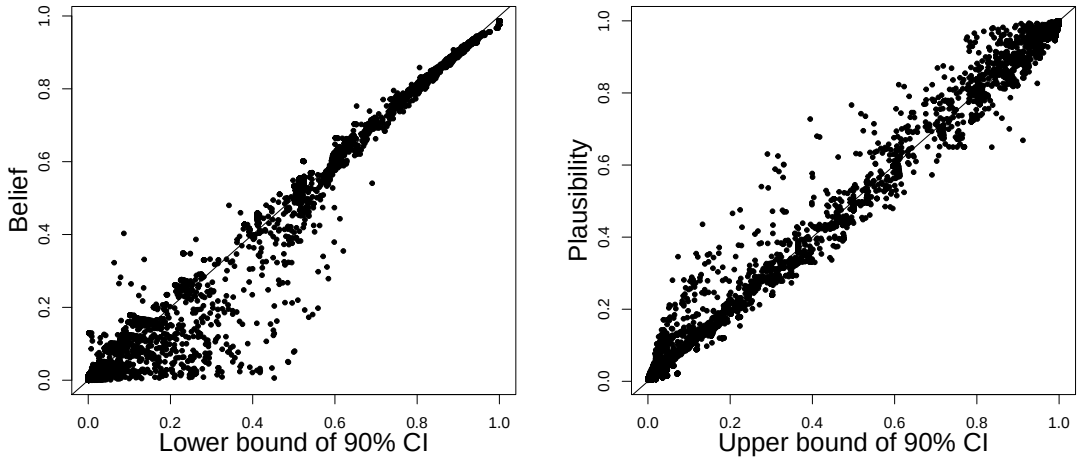


Figure 19: Diabetes data with the partition obtained by fitting a GMM with $c = 3$ components. Covariances in each group are represented by isodensity ellipses. (This figure is better viewed in color).



(a) (b)

Figure 20: Approximation of confidence intervals by belief-plausibility intervals for the Diabetes data. (a) Lower bound P_{ij}^l of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. belief degree $Bel_{ij}(\{s_{ij}\})$ (y-axis); (b) Upper bound P_{ij}^u of the 90% confidence interval on $P_{ij}(\boldsymbol{\theta})$ (x-axis) vs. plausibility degree $Pl_{ij}(\{s_{ij}\})$ (y-axis).

Table 10: Confusion matrix for the evidential partition of the Diabetes dataset (model-based approach).

	Clustering				
	{ch}	{no}	{ov}	{ch, no}	{no, ov}
chemical	18	6	0	8	4
normal	2	68	0	6	0
overt	2	0	25	0	6

and the quality of the approximation of confidence intervals by belief-plausibility intervals is illustrated in Figure 20. The confusion matrix is shown in Table 10. As we can see, the number of misclassifications is down to 14 (including 4 objects of class "chemical" wrongly assigned to {no, ov}). As a comparison, we show the confusion matrices for ECM (Table 11) and k-EVCLUS (Table 12), which were used with the same parameter settings as for the Iris data. As we can see, these two methods fail to group the observations from the class "chemical" in a single cluster, and they perform significantly worse than the model-based approach.

GvHD data

The GvHD (Graft-versus-Host Disease) data[4] consist of four biomarker variables, namely, CD4, CD8b, CD3, and CD8, observed in flow cytometry data for two patients [6, 43]. We

[4]Available in the R package `mclust`.

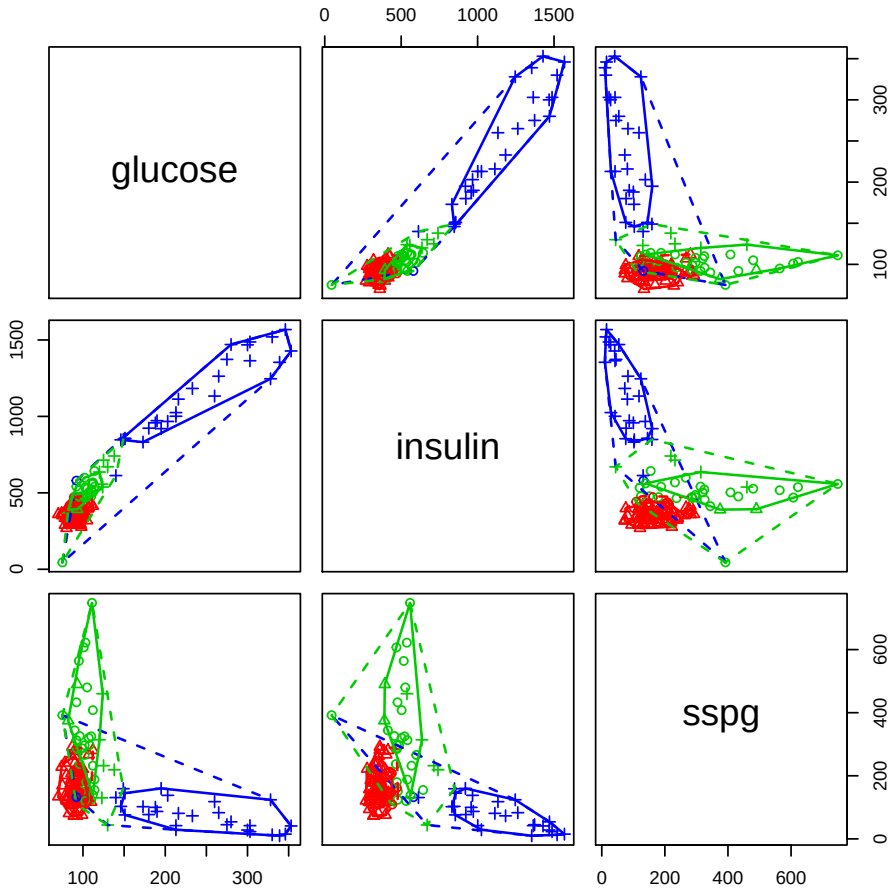


Figure 21: Evidential partition of the Diabetes data obtained using the model-based approach. The true groups are represented by different symbols (o: chemical; triangle: normal; +: overt), and the maximum-plausibility groups are represented by different colors. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

Table 11: Confusion matrix for the evidential partition of the Diabetes dataset obtained by ECM.

	Clustering			
	{ch}	{no}	{ov}	{ch, no}
chemical	8	17	0	11
normal	2	68	0	6
overt	1	10	21	1

Table 12: Confusion matrix for the evidential partition of the Diabetes dataset obtained by k-EVCLUS.

	Clustering				
	{ch}	{no}	{ov}	{no, ov}	{ch, no, ov}
chemical	8	27	0	0	1
normal	1	75	0	0	0
overt	1	9	21	2	0

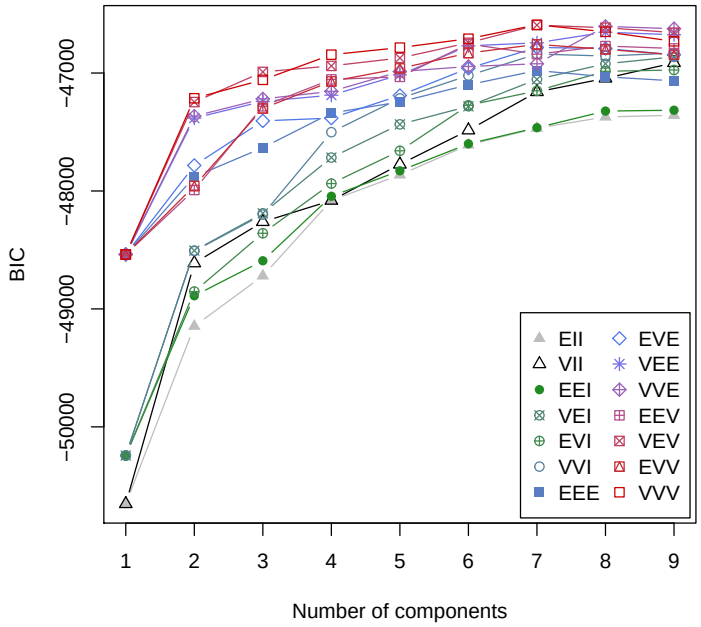


Figure 22: Model selection for the GvHD: BIC vs. number of clusters for the 14 models defined in R package `mclust`.

used the data from the GvHD positive patient, which originally contained 9083 observations. We randomly selected 1000 observations. The objective of the analysis is to identify cell sub-populations present in the sample. There are no ground truth labels for this dataset, but we use it to as an example of a dataset with a larger number of clusters than the two previous datasets.

As seen in Figure 22, the best model according to BIC is the full (unconstrained) model with $c = 7$ clusters. The corresponding partition as well as the cluster centers and covariance ellispses are shown in Figure 23.

With seven clusters, the maximum number of nonempty focal sets in the evidential partition is $2^7 - 1 = 127$. Restricting the focal sets to singletons and pairs leaves us with $7 + (6 \times 7)/2 = 28$ focal sets. However, not all pairs are needed, because some pairs of clusters actually do not overlap. To further reduce the number of focal sets, we can use a

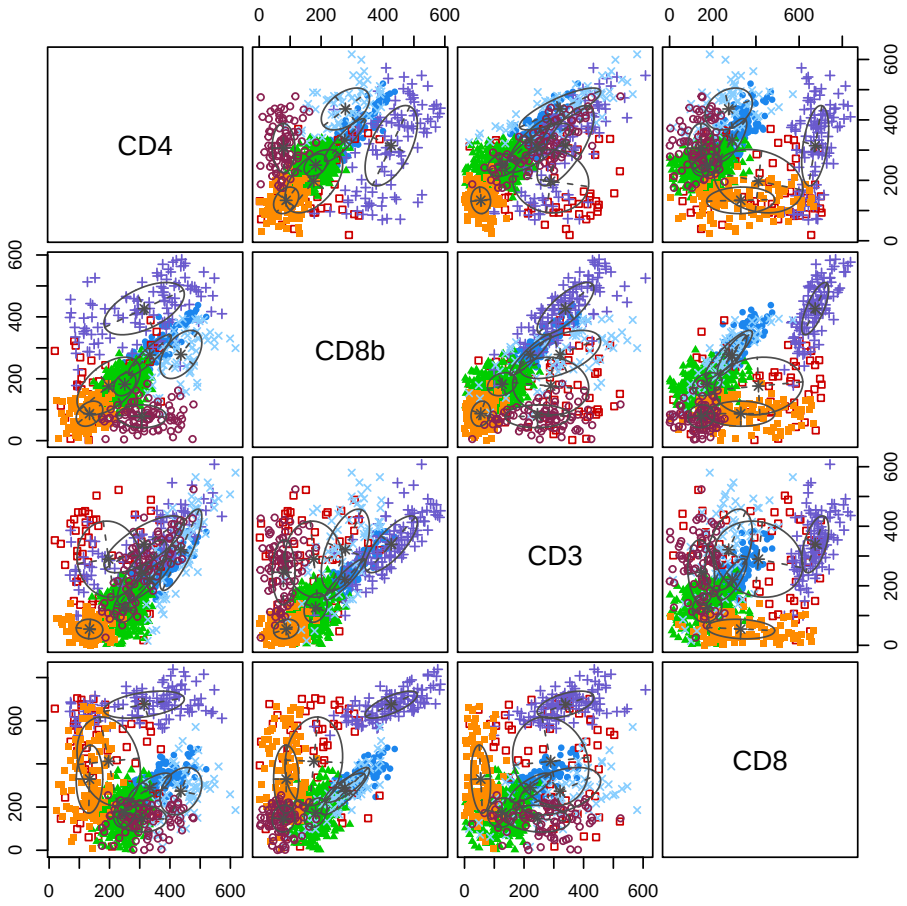


Figure 23: GvHD data with the partition obtained by fitting a GMM with $c = 7$ components. Covariances in each group are represented by isodensity ellipses. (This figure is better viewed in color).

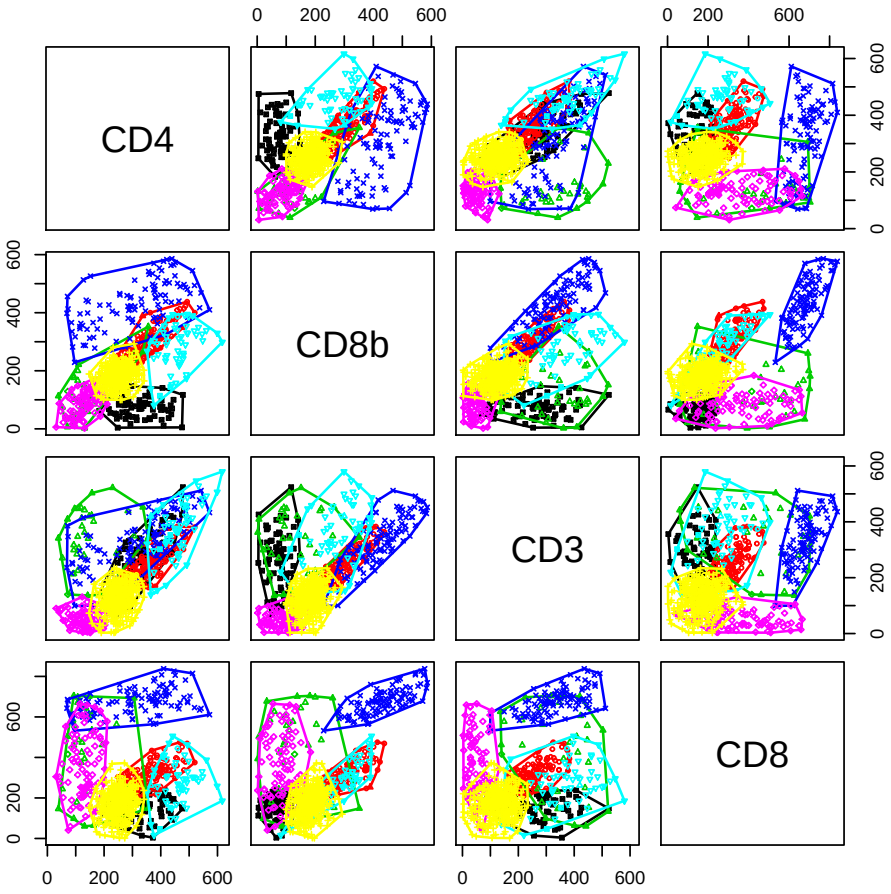


Figure 24: Evidential partition of the GvHD data: lower approximations and convex hulls of the upper approximations. The solid and broken lines represent, respectively, the convex hulls of the lower and upper approximation of each cluster. (This figure is better viewed in color).

method similar to the one proposed in [14]. The similarity between two clusters k and l can be measured by

$$s_{kl} = \sum_{i=1}^{n} \pi_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}})\, \pi_l(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}).$$

Based on these similarities, we can identify clusters that are mutual K-nearest neighbors. With $K = 2$, we obtained five pairs of mutual nearest neighbors: (1,3), (2,4), (1,6), (3,7) and (5,7). These five pairs and the seven singletons gave us $f = 12$ focal sets. We used the same method as above to compute the bootstrap percentile confidence intervals and construct an evidential partition. The cluster lower approximations and the convex hulls of the upper approximations are shown in Figure 24.

Using this pair selection approach, the model can be used even with large numbers of clusters (several dozens or even several hundreds). The main limitation of the method is

related to the number of objects. The necessity to compute and store the $n(n-1)/2$ belief-plausibility intervals results in a quadratic memory and time complexity, which precludes application of the method to datasets with more than a few thousand objects. However, it might be possible to use only pairwise belief-plausibility intervals for pairs of neighboring objects, as done in [14] to make the EVCLUS algorithm applicable to large datasets. This idea remains to be investigated.

5. Conclusions

We have described a new model-based approach to evidential clustering. The method starts by estimating the parameters of a finite mixture model. In this paper, we used GMMs, but there is no restriction on the kinds of models that can be used. For instance, for categorical data, latent class models would be more suitable. The model is first fitted using the EM algorithm, and bootstrap percentile confidence intervals on pairwise probabilities P_{ij} at some confidence level $1-\alpha$ are computed. Here, P_{ij} is the probability that objects i and j belong to the same cluster. Finally, an evidential partition is constructed in such a way that pairwise degrees of belief $Bel_{ij}(\{s_{ij}\})$ and plausibility $Pl_{ij}(\{s_{ij}\})$ approximate the bounds of the confidence intervals in the least squares sense. The evidential partitions constructed using this method are approximately calibrated, in the sense that the belief-plausibility intervals $[Bel_{ij}(\{s_{ij}\}), Pl_{ij}(\{s_{ij}\})]$ contain the true probabilities P_{ij} with probability approximately equal to $1-\alpha$. This evidential partition provides a more complete description of the clustering structure than does the fuzzy partition directly provided by the EM algorithm, as it also takes into account uncertainty in the estimation of class probabilities.

We have presented extensive experimental results showing that the coverage probabilities of the belief-plausibility intervals are close to their nominal confidence level when the model is correctly specified. We have also demonstrated the applicability of this approach to several real datasets, and compared the evidential partitions obtained using this model-based approach to those obtained with ECM and EVCLUS, the two main evidential clustering algorithms available so far. Model-based evidential clustering inherits the advantages of classical model-based clustering. In particular, various assumptions about cluster shapes can be formalized as assumptions about component probability distributions, and model selection criteria such as BIC make it possible to determine the number of clusters automatically.

As the method requires the construction of confidence intervals for each pair objects, it has quadratic complexity, which makes it unsuitable for the analysis of very large datasets containing more than a few thousands of objects. One remedy could be to use only the belief-plausibility intervals for pairs of neighboring objects, an idea exploited in [14] to apply the EVCLUS algorithm to large datasets. This research direction will be explored in future work.

References

[1] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49: 803–821, 1993.

[2] Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.

[3] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*. Kluwer Academic Publishers, Boston, 1999.

[4] J.C Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, New-York, 1981.

[5] Manuel Blum, Robert W. Floyd, Vaughan Pratt, Ronald L. Rivest, and Robert E. Tarjan. Time bounds for selection. *Journal of Computer and System Sciences*, 7(4):448–461, 1973.

[6] R. R. Brinkman, M. Gasparetto, S.-J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Biology of Blood and Marrow Transplantation*, 13:691–700, 2007.

[7] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5): 781–793, 1995.

[8] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, New-York, 1997.

[9] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[11] T. Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.

[12] T. Denoeux and O. Kanjanatarakul. Beyond fuzzy, possibilistic and rough: An investigation of belief functions in clustering. In *Soft Methods for Data Science (Proc. of the 8th International Conference on Soft Methods in Probability and Statistics SMPS 2016)*, volume AISC 456 of *Advances in Intelligent and Soft Computing*, pages 157–164, Rome, Italy, September 2016. Springer-Verlag.

[13] T. Denœux and M.-H. Masson. EVCLUS: Evidential clustering of proximity data. *IEEE Trans. on Systems, Man and Cybernetics B*, 34(1):95–109, 2004.

[14] T. Denœux, S. Sriboonchitta, and O. Kanjanatarakul. Evidential clustering of large dissimilarity data. *Knowledge-based Systems*, 106:179–195, 2016.

[15] Thierry Denoeux. Decision-making with belief functions: a review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.

[16] Thierry Denœux. *evclust: Evidential Clustering*, 2020. URL `https://CRAN.R-project.org/package=evclust`. R package version 1.1.0.

[17] Thierry Denœux and Shoumei Li. Frequency-calibrated belief functions: Review and new insights. *International Journal of Approximate Reasoning*, 92:232–254, 2018.

[18] Thierry Denoeux, Shoumei Li, and Songsak Sriboonchitta. Evaluating and comparing soft partitions: an approach based on Dempster-Shafer theory. *IEEE Transactions on Fuzzy Systems*, 26(3):1231–1244, 2018.

[19] Thierry Denœux, Didier Dubois, and Henri Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, chapter 4. Springer Verlag, 2020.

[20] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3): 189–212, 1996.

[21] Pierpaolo D'Urso. Informational paradigm, management of uncertainty and theoretical formalisms in the clustering framework: A review. *Information Sciences*, 400—401:30–62, 2017.

[22] Pierpaolo D'Urso and Riccardo Massari. Fuzzy clustering of mixed data. *Information Sciences*, 505: 513–534, 2019.

[23] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New-York, 1993.

[24] Alessio Ferone and Antonio Maratea. Integrating rough set principles in the graded possibilistic clustering. *Information Sciences*, 477:148–160, 2019.

[25] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3618, 1990.

[26] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Englewood Cliffs, NJ., 1988.
[27] R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1:98–111, 1993.
[28] Benoît Lelandais, Su Ruan, Thierry Denœux, Pierre Vera, and Isabelle Gardin. Fusion of multi-tracer PET images for dose painting. *Medical Image Analysis*, 18(7):1247–1259, 2014.
[29] Feng Li, Shoumei Li, and Thierry Denœux. k-CEVCLUS: Constrained evidential clustering of large dissimilarity data. *Knowledge-Based Systems*, 142:29–44, 2018.
[30] C. Lian, S. Ruan, T. Denoeux, H. Li, and P. Vera. Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images. *IEEE Transactions on Biomedical Engineering*, 65(1):21–30, 2018.
[31] Pawan Lingras and Georg Peters. Applying rough set concepts to clustering. In G. Peters, P. Lingras, D. Ślezak, and Y. Yao, editors, *Rough Sets: Selected Methods and Applications in Management and Engineering*, pages 23–37. Springer-Verlag, London, UK, 2012.
[32] Nasr Makni, Nacim Betrouni, and Olivier Colot. Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: Application to prostate multi-parametric MRI. *Information Fusion*, 19:61–72, 2014.
[33] M.-H. Masson and T. Denoeux. ECM: an evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008.
[34] M.-H. Masson and T. Denœux. RECM: relational evidential c-means algorithm. *Pattern Recognition Letters*, 30:1015–1026, 2009.
[35] G. J. McLachlan and K. E. Basford. *Mixture Models: inference and applications to clustering*. Marcel Dekker, New York, 1988.
[36] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.
[37] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.
[38] Adrian O'Hagan, Thomas Brendan Murphy, Luca Scrucca, and Isobel Claire Gormley. Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, 34:1779–1813, 2019.
[39] Georg Peters. Rough clustering utilizing the principle of indifference. *Information Sciences*, 277:358 – 374, 2014.
[40] Georg Peters. Is there any need for rough clustering? *Pattern Recognition Letters*, 53:31–37, 2015.
[41] Georg Peters, Fernando Crespo, Pawan Lingras, and Richard Weber. Soft clustering: fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*, 54 (2):307–322, 2013.
[42] G. M. Reaven and R. G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16:17–24, 1979.
[43] Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016. URL `https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf`.
[44] Lisa Serir, Emmanuel Ramasso, and Noureddine Zerhouni. Evidential evolving Gustafson-Kessel algorithm for online data streams partitioning using belief function theory. *International Journal of Approximate Reasoning*, 53(5):747–768, 2012.
[45] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
[46] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer, New-York, 1995.
[47] Cajo J.F. ter Braak, Yiannis Kourmpetis, Henk A.L. Kiers, and Marco C.A.M. Bink. Approximating a similarity matrix by a latent class model: A reappraisal of additive fuzzy clustering. *Computational Statistics & Data Analysis*, 53(8):3183–3193, 2009.
[48] Stephen A. Vavasis. Complexity theory: quadratic programming. In Christodoulos A. Floudas and Panos M. Pardalos, editors, *Encyclopedia of Optimization*, pages 304–307. Springer US, Boston, MA, 2001.
[49] K. Wang, S. Ng, and G. J. McLachlan. Multivariate skew t mixture models: Applications to fluorescence-activated cell sorting data. In *2009 Digital Image Computing: Techniques and Appli-*

cations, pages 526–531, Dec 2009. doi: 10.1109/DICTA.2009.88.

[50] Kui Wang, Angus Ng, and Geoff McLachlan. *EMMIXskew: The EM Algorithm and Skew Mixture Distribution*, 2018. URL `https://CRAN.R-project.org/package=EMMIXskew`. R package version 1.0.3.

[51] Miin-Shen Yang, Shou-Jen Chang-Chien, and Yessica Nataliani. Unsupervised fuzzy model-based gaussian clustering. *Information Sciences*, 481:1–23, 2019.

[52] Kuang Zhou, Arnaud Martin, Quan Pan, and Zhun-Ga Liu. Median evidential c-means algorithm and its application to community detection. *Knowledge-Based Systems*, 74(0):69–88, 2015.