Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

Christophe Biernacki, Matthieu Marbac, Vincent Vandewalle

Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

C. Biernacki
Inria Lille and University Lille 1, CNRS
christophe.biernacki@inria.fr

M. Marbac
CREST and ENSAI
matthieu.marbac-lourdelle@ensai.fr

V. Vandewalle
EA 2694 University Lille 2
vincent.vandewalle@inria.fr

December 10, 2019

Abstract

A generic method is introduced to visualize in a "Gaussian-like way", and onto $\mathbb{R}^2$, results of Gaussian or non-Gaussian based clustering. The key point is to explicitly force a visualization based on a spherical Gaussian mixture to inherit from the within cluster overlap that is present in the initial clustering mixture. The result is a particularly user-friendly drawing of the clusters, providing any practitioner with an overview of the potentially complex clustering result. An entropic measure provides information about the quality of the drawn overlap compared to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis.

Keywords: Dimension reduction, Gaussian mixture, factorial analysis, linear discriminant analysis, model-based clustering, visualization.

Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

Abstract

A generic method is introduced to visualize in a "Gaussian-like way", and onto $\mathbb{R}^2$, results of Gaussian or non-Gaussian based clustering. The key point is to explicitly force a visualization based on a spherical Gaussian mixture to inherit from the within cluster overlap that is present in the initial clustering mixture. The result is a particularly user-friendly drawing of the clusters, providing any practitioner with an overview of the potentially complex clustering result. An entropic measure provides information about the quality of the drawn overlap compared to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis.

Keywords: Dimension reduction, Gaussian mixture, factorial analysis, linear discriminant analysis, model-based clustering, visualization.

1 Introduction

Data analysis is the exploratory field of multivariate statistics. It essentially encompasses the clustering and the visualization tasks. Both are often jointly involved: either visualization is performed in the hope of revealing the "graphical evidence" of a cluster structure in the data set; or clustering is performed first and the visualization task follows in the hope of providing a better understanding of the estimated cluster structure. We are primarily interested in the second scenario.

Clustering (Jajuga et al. 2002) serves to summarize (typically large) data sets by assessing a partition among observations, the latter being thus summarized by (typically few) characteristic classes. Model-based clustering (McLachlan & Peel 2004, McNicholas 2016, Biernacki 2017) achieves the clustering purpose in a probabilistic framework, usually

consisting of modeling the whole data distribution using a finite mixture model. Classical challenges can thereby be solved by using tools that rely on theoretical statistics, *e.g.* estimating the partition using an EM algorithm (Dempster et al. 1977), selecting the number of groups using information criteria such as BIC or ICL (Schwarz 1978, Biernacki et al. 2000), dealing with missing values among observations (Larose 2015). Moreover, this framework allows for the analysis of different types of data by "simply" adapting the related cluster distribution: continuous data (Banfield & Raftery 1993, Celeux & Govaert 1995, McNicholas & Murphy 2008), categorical data (Goodman 1974, Celeux & Govaert 1991, Gollini & Murphy 2014, Marbac et al. 2016), mixed data (Kosmidis & Karlis 2015, McParland & Gormley 2016, Punzo & Ingrassia 2016, Marbac et al. 2017, Mazo 2017), functional data (Samé et al. 2011, Bouveyron & Jacques 2011, Jacques & Preda 2014), networks data (Daudin et al. 2008, Zanghi et al. 2008, Ambroise & Matias 2012).

Once the clustering process has been performed, the next step is to provide a good understanding of it to practitioners. However, a rendering based on a raw delivery of the model parameters and/or the resulting partition (or the related conditional membership probabilities) can be quite inefficient: understanding of the parameters requires specific knowledge of the model at hand the partition can be also hard to read since it is just a numerical list the length of the sample size, which must be large enough to have initially motivated the clustering process.

Visualization is designed to express, in a user-friendly manner, the estimated clustering structure. Its general principle is to design a mapping of the data, or of other related statistical results such as the cluster shape, within a "friendly" space (generally $\mathbb{R}^2$) while maintaining some properties that the data, or the related statistical results, have in their native space. The vast majority of proposed mapping relies on different variants of factorial analysis or other distance-based methods (like multidimensional scaling). For a thorough

list of visualization methods, see Section 2.2, and references therein. However, all standard mappings waste most clustering information that is conveyed by the probabilistic approach, except Scrucca (2010) which uses the full model-based approach for the mapping. However, this approach is limited to continuous data.

This paper defends the key idea that only a so-called model-based visualization output can exploit the model-based clustering input, since both involved objects are of the same nature (probabilistic objects). More precisely, the mixture model used for the visualization output will inherit from the overlap of the initial mixture model. In fact, this is similar to defining a particular mapping but without any explicit distance design. This process has the clear advantage of being straightforwardly suitable for any type of data without any specific definition of the mixture output since only the conditional memberships need to be estimated. In fact, the specificity of initial data has been taken into account by the initial clustering modeling process. The mixture output involves spherical Gaussian components, with the same number of components as the clustering mixture. This particular Gaussian choice is informed by both some technical arguments and some user-friendly arguments. The resulting drawing displays meaningful spherical cluster shapes in the bivariate continuous space. Finally, accuracy of this drawing is assessed by comparing the apparent overlap mixture on the graph and the overlap of the initial mixture. To have a good understanding of our proposal, in particular its link with model-based clustering techniques, its general outline can be summarized as follows:

1. select a model-based clustering technique for data at hand;
2. extract the whole distribution fo the classification probabilities from the fitted model;
3. fit a multivariate spherical Gaussian mixture respecting as far as possible the distribution of the previous classification probabilities;

4. (a) draw the spherical Gaussian mixture pdf on the most discriminative bivariate map;

   (b) draw a "pseudo" bivariate scatter plot representing the individual classification probabilities on the most discriminative bivariate map.

This paper is organized as follows. Section 2 focuses on the context of model-based clustering for mixed data and reviews the main existing visualization techniques of a clustering result. Section 3 presents the central contribution of this work consisting of matching any clustering mixture and a spherical multivariate Gaussian visualization mixture according to their component overlap, and then describes in Section 4 how to draw this Gaussian mixture in the most discriminative map. Like any visualization method, our proposition can introduce a bias. However, because we propose a full model-based visualization approach, an index measuring this bias (and thus the quality of the representation) is presented. Section 5 then proposes a means of displaying a kind of individual plotting on the same discriminative map to access each individual data cluster membership positioning. Section 6 illustrates in depth the Gaussian model-based proposition on three real data sets with different types of features (mixed data, functional data and network data). Throughout this paper, the proposition is also illustrated via a categorical running example. Section 7 concludes this work.

2 Clustering: from modeling to visualizing

2.1 Model-based clustering of multi-type data

Clustering aims to estimate a partition $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, composed of K clusters, of a data set $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, composed of n observations. The component membership of each

observation $\boldsymbol{x}_i$ is given by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$, with $z_{ik} = 1$ if $\boldsymbol{x}_i$ arises from component k and $z_{ik} = 0$ otherwise. $\mathcal{Z}$ denotes the space of any $\boldsymbol{z}_i$. In a very general situation, each observation $\boldsymbol{x}_i$ is defined on a space $\mathcal{X}$ described by d_X variables which can be continuous, categorical or functional.

Model-based clustering aims to solve the clustering task in a full probabilistic framework by modeling the distribution of the full data set $(\mathbf{x}, \mathbf{z})$, $\mathbf{z}$ being considered as a latent part of the data set. This framework has the decisive advantage of consolidating the exploratory clustering result through the background of mathematical statistics (estimation, model selection; see for instance McLachlan & Peel (2004), Biernacki (2017)). More precisely, all couples $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ are assumed to independently arise from the distribution defined by the probability density function (pdf)

$$f(\boldsymbol{x}_i, \boldsymbol{z}_i) = \prod_{k=1}^{K} \left[\pi_k f_k(\boldsymbol{x}_i)\right]^{z_{ik}} \tag{1}$$

where π_k is the proportion of the kth component ($\pi_k > 0$ and $\sum_k \pi_k = 1$) and f_k is the pdf of this component.

From such a modeling, two interesting by-product distributions are available. Firstly, the (marginal) distribution of each $\boldsymbol{x}_i$ corresponds to the so-called K component mixture defined by the pdf $f(\boldsymbol{x}_i) = \sum_k \pi_k f_k(\boldsymbol{x}_i)$. Secondly, under distribution f, the probability that $\boldsymbol{x}_i$ arises from component k, denoted by $t_{ik}(f)$, is expressed by

$$t_{ik}(f) = \mathrm{p}(z_{ik} = 1 | \boldsymbol{x}_i; f) = \frac{\pi_k f_k(\boldsymbol{x}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x}_i)}. \tag{2}$$

Thus, all information about the classification probabilities for observation $\boldsymbol{x}_i$ can be stored in a $K-1$ continuous vector $\boldsymbol{t}_i(f)$ (because $\sum_{k=1}^{K} t_{ik}(f) = 1$) where

$$\boldsymbol{t}_i(f) = (t_{i1}(f), \ldots, t_{i(K-1)}(f)). \tag{3}$$

Information about the classification probabilities of sample $\mathbf{x}$ is given by $\mathbf{t}(f) = (\boldsymbol{t}_1(f), \ldots, \boldsymbol{t}_n(f))$.

Traditionally, components f_k are parametrized by finite dimensional vectors and an EM algorithm, or one of its variants (Dempster et al. 1977), is used to provide an estimate $\hat{f}$ of f (the π_k's and the parameters associated with the f_k's). Alternatively semi- or non-parametric mixtures can be considered (Benaglia et al. 2009). Finally, an estimated partition $\hat{\mathbf{z}}$ can be straightforwardly deduced from $\mathbf{t}(\hat{f})$ by using the rule of maximum *a posteriori* (MAP) defined by $\hat{z}_{ik} = 1$ iif $k = \arg\max_\ell t_{i\ell}(\hat{f})$.

Thus, the key point to achieve this model-based clustering procedure is to define the distributional space $\mathcal{F}$ where f stands for $(f \in \mathcal{F})$. In fact, only the space of components f_k has to be defined. Clearly, choosing component pdf f_k depends on $\mathcal{X}$. Many proposals already exist such as multivariate Gaussian or multivariate t-distributions for continuous data (McLachlan & Peel 2004, McNicholas 2016), a product of multinomial distributions for categorical data (Goodman (1974); see also the running example later), a product of Gaussian and multinomial distributions when mixing continuous and categorical data (Moustaki & Papageorgiou 2005, see also numerical experiments in Section 6.1), specific models for functional data or for network data (see respectively numerical experiments in Section 6.2 and 6.3, with references therein).

However, because of their potential complexity, such previous mathematical features may fail to provide a *user-friendly* clustering understanding. Indeed, it may be difficult to have a useful overview of individuals in clusters through $\hat{\mathbf{z}}$ (or through $\mathbf{t}(\hat{f})$) if n or K is too large. Similarly, it may be difficult to get a useful overview of the whole clusters (proportions, shapes, positioning, *etc.*) through $\hat{f}$ if the space $\mathcal{X}$ involves many features (d large) or involves features of complex types (like a mix of categorical and functional features), a situation where the pdf of the components can be particularly hard to embrace as a whole. As a matter of fact, the need for a user-friendly understanding of the math-

ematical clustering results (at both individual and pdf levels) is the very reason for using some specific visualization procedures.

2.2 Overview of clustering visualization

Mapping *vs.* **drawing** Visualization is probably one of the most appealing data analysis tasks for practitioners since its fundamental purpose is to display some potentially complex and technically demanding statistical objects (typically a data set or a pdf) on simple and seamlessly accessible graphs (typically a scatter plot or an iso-density curve). The whole process can be viewed as the achievement of two different successive steps. The *mapping step* transforms the initial statistical object into a simpler statistical one typically through a space dimension reduction of a data set or of a pdf (marginal pdf). It produces no graphical output at all. The *drawing step* provides the final graphical display from the output of the previous mapping step and usually entails the use of conventional graphical toolboxes. It fine-tunes *all* the possible graphical parameters.

Individual mapping The clustering visualization task is probably thought as firstly as visualizing simultaneously the data set $\mathbf{x}$ and its estimated partition $\hat{\mathbf{z}}$. Typically, the corresponding mapping, designated below by M^{ind}, transforms the data set $\mathbf{x}$, defined on $\mathcal{X}$, into a new data set $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, defined on a new space $\mathcal{Y}$, as follows:

$$M^{\text{ind}} \in \mathcal{M}^{\text{ind}} : \ \mathbf{x} \in \mathcal{X}^n \mapsto \mathbf{y} = M^{\text{ind}}(\mathbf{x}) \in \mathcal{Y}^n. \tag{4}$$

Here $\mathcal{M}^{\text{ind}}$ denotes a particular mapping family. This family varies according to the type of data involved in $\mathcal{X}$ and also depending on whether they use only data $\mathbf{x}$ or additional clustering information $\hat{\mathbf{z}}$ or $\mathbf{t}(\hat{f})$.

Methods relying on data $\mathbf{x}$ (thus discarding clustering information) are certainly the

most frequent. In terms of continuous data, *principal component analysis* (PCA; Josse et al. (2011), Verbanck et al. (2015), Audigier et al. (2016*a*)) serves to represent the data on a map by focusing on their dispersion. Similarly, categorical data can be visualized using *multiple correspondence analysis* (MCA; Van der Heijden & Escofier (2003), Josse et al. (2012), Greenacre (2017)), a mix of continuous and categorical data can be visualized using *mixed factorial analysis* (MFA; Chavent et al. (2012), Audigier et al. (2016*b*)) and functional data can be visualized using *functional principal component analysis* (FPCA; Ramsay & Silverman (2005), Zhou & Pan (2014), Chen & Lei (2015)). *Multidimensional scaling* (MDS; Young (1987), Cox & Cox (2001)) is more general since it can be used to deal with any type of data. It relies on dissimilarities between pairs of individuals for inputs $\mathbf{x}$ and also for outputs $\mathbf{y}$, the resulting coordinate matrix $\hat{\mathbf{y}}$ being obtained by minimizing a loss function. However, dissimilarities have to be defined specifically in respect of the type of data under consideration. For just illustrating this point, the Euclidean distance is frequent for continuous data whereas the Hamming distance is more suitably for binary data.

In an machine learning framework, methods such as *self-organized map* (SOM; Kohonen (1982)) or generative topographic mapping (GTM; Bishop et al. (1998)) have been developed to summarize the data in terms of a set of reference points having a regular spatial organization corresponding generally to a two-dimensional regular network. But, even if nodes of the network are usually interpreted as clusters, these ones essentially serve as a preprocessing step for limiting the number of prototypes to be considered at a second step in a hierarchical clustering (Vesanto & Alhoniemi 2000).

Methods taking into account additional clustering information $\hat{\mathbf{z}}$ or $\mathbf{t}(\hat{f})$ are less common and are mostly restricted to continuous data. We can cite *linear discriminant analysis* (LDA; Fisher (1936), Xanthopoulos et al. (2013)) which takes into account cluster sep-

aration by defining the mapping through to a particular factorial analysis of the cluster means. Also, in the specific case of continuous data, Hennig (2004), Scrucca (2010) and Morris et al. (2013) defined a specific linear mapping between $\mathcal{X}$ and $\mathcal{Y}$. In that case, the distribution of $\mathbf{y}$ is itself a (less-dimensional) Gaussian mixture or a multivariate t-mixture, with the same number of components and the same proportions, which can be expressed as $g = \sum_k \pi_k g_k$. Finally, their method aims to preserve the related conditional membership probabilities $\mathbf{t}(\hat{f})$ and $\mathbf{t}(g)$, namely the classification probabilities of $\mathbf{x}$ with $\hat{f}$ and the classification probabilities of $\mathbf{y}$ with g, respectively. In other words, the aim is to find a linear mapping that preserves as far as possible, through the mapping mixture g, the cluster separation occurring in the original mixture f. Somewhat the method we proposed in this paper is related to this idea but it is not restricted to continuous distributions in the mixture and it does not relay on a linear mapping.

Pdf mapping Many visualizations are in practice overlaid by additional information relating to the corresponding mapping distribution. This mapping transforms the initial mixture $f = \sum_k \pi_k f_k$, defined on the distributional space $\mathcal{F}$, into a new mixture $g = \sum_k \pi_k g_k$, defined on the distributional space $\mathcal{G}$. It can be expressed as the following mapping, designated here by M^{pdf}:

$$M^{\text{pdf}} \in \mathcal{M}^{\text{pdf}} : \; f \in \mathcal{F} \mapsto g = M^{\text{pdf}}(f) \in \mathcal{G}, \tag{5}$$

where $\mathcal{M}^{\text{pdf}}$ denotes a particular mapping family. It is important to note that the pdf mapping M^{pdf} is rarely defined "from scratch" since it can be obtained as a "simple" by-product from the previous individual mapping M^{ind}. However, in practice, the resulting mixture g can be particularly tedious to calculate (possibly no closed-form solution available outside linear mappings), which can be partially overcome by displaying the empirical mapping of a very large sample. But the resulting pdf can also have non-conventional

iso-density shape per cluster (for instance clusters with disconnected parts), undermining somewhat all the user-friendliness that is expected when using pdf visualization.

2.3 Running example

As a running example for this paper, we consider the data set of Schlimmer (1987). It is composed of votes for each of the $n = 435$ U.S. House of Representatives Congressmen on $d_X = 16$ key votes. For each vote, three levels are considered: yea, nay or unknown disposition. Data are clustered by a mixture of products of multinomial distributions (Goodman 1974). Parameter estimation is performed by maximum likelihood and model selection is done by the BIC (Schwarz 1978), which selects $K = 4$ components. The R package Rmixmod (Lebret et al. 2015) is used for inference.

As an output of this estimation step, the user is provided with a partition and a parameter. It may be not really convenient to have a detailed look at the partition of 435 individuals. In regard to the parameters, the mixing proportions can be suitable for a quick, but partial, understanding of the clustering result. However, going further into the clustering understanding by analyzing the multinomial parameters can be very laborious since it entails $192 = 16 \times 3 \times 4$ values to be observed and compared.

It is also possible to analyze the clustering results graphically in a conventional way. Figure 1 presents the scatter plot of the Congressmen and their partition on the first map of the MCA, obtained by the R package FactoMineR (Lê et al. 2008). It appears that the scatter plot provided by MCA is quite hard to read. Firstly, it is well-known that total inertia is hard to interpret, and consequently the information about a possible relative positioning of clusters can be questionable. Secondly, even if faithful, overlap between components is not fully visible and thus does not allow for a straightforward interpretation of f.

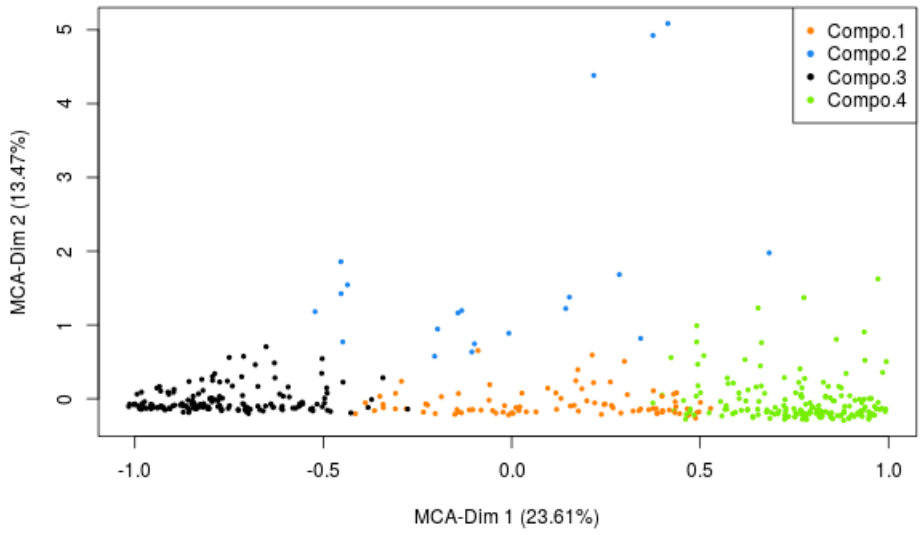


Figure 1: Scatter plot of the Congressmen and their partition on the first MCA map.

3 Mapping clusters as spherical Gaussians

In this section, we focus our attention on the so-called pdf visualization, arguing that, asymptotically on the sample size n, similar objects result from both pdf and individual visualization processes. However, we will hold a specific discussion on individual visualization below.

3.1 Changing the mapping objects to be controlled

Traditional way: controlling the mapping family As described in Section 2.2, the cornerstone of all traditional pdf visualization procedures is based on defining the mapping family $\mathcal{M}^{\text{pdf}}$ (or more exactly $\mathcal{M}^{\text{ind}}$ from which $\mathcal{M}^{\text{pdf}}$ is almost always deduced). As just

an example, the reader can have in mind the classical linear mapping for the continuous case. Then, the pdf family $\mathcal{G}$ of g is a simple by-product of $\mathcal{M}^{\text{pdf}}$, and thus can be denoted by $\mathcal{G}(\mathcal{M}^{\text{pdf}})$. Using the general mapping expression (5), $\mathcal{G}(\mathcal{M}^{\text{pdf}})$ is naturally expressed as follows:

$$\mathcal{G}(\mathcal{M}^{\text{pdf}}) = \{g : g = M^{\text{pdf}}(f), f \in \mathcal{F}, M^{\text{pdf}} \in \mathcal{M}^{\text{pdf}}\}. \tag{6}$$

As an immediate consequence, the nature of $\mathcal{G}$ can depend to a great extent on the choice of $\mathcal{M}^{\text{pdf}}$, leading potentially to very different cluster shapes. Arguments that lead to traditional $\mathcal{M}^{\text{pdf}}$ (or $\mathcal{M}^{\text{ind}}$) rely essentially on a combination of user-friendly and easy-to-compute properties. For instance, in the continuous case, linear mappings are often retained (like for PCA). In the categorical case, a continuous space $\mathcal{Y}$ is often targeted (like for MCA). It is a similar situation for functional data with FPCA or also for mixed data with MFA or MDS, even if MDS is a somewhat more complex procedure since it is not always defined in closed-form. However, such choices may vary significantly from one statistician to another one. For instance, MDS relies on defining dissimilarities both inside spaces $\mathcal{X}$ and $\mathcal{Y}$ and changing them could significantly affect the resulting mapping.

New proposed method: controlling the distribution family Alternatively, the general mapping expression (5) can be seen as indexed by the distribution family $\mathcal{G}$, the mapping $\mathcal{M}^{\text{pdf}}$ being now obtained as a by-product, and thus now denoted by $\mathcal{M}^{\text{pdf}}(\mathcal{G})$. This new point of view is straightforwardly expressed as:

$$\mathcal{M}^{\text{pdf}}(\mathcal{G}) = \{M^{\text{pdf}} : g = M^{\text{pdf}}(f), f \in \mathcal{F}, g \in \mathcal{G}\}. \tag{7}$$

It corresponds to the reversed situation of (6) where $\mathcal{G}$ has to be defined instead of M^{pdf}. This new freedom indeed provides an opportunity to directly force $\mathcal{G}$ to be a user-friendly mixture family.

3.2 Constrained spherical Gaussians as matching candidates

Spherical Gaussians One of the most simple and natural candidate belonging to the "user-friendly mixture family" is probably the spherical Gaussian mixture defined on $\mathcal{Y} = \mathbb{R}^{d_Y}$. Its pdf is defined for any $\boldsymbol{y} \in \mathbb{R}^{d_Y}$ by

$$g(\boldsymbol{y}; \boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k \phi_{d_Y}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{I}), \tag{8}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\phi_{d_Y}(.; \boldsymbol{\mu}_k, \boldsymbol{I})$ is the pdf of the Gaussian distribution with mean $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kd_Y}) \in \mathbb{R}^{d_Y}$ and covariance matrix equal to identity $\boldsymbol{I}$.

Because clustering visualization is the central task of this work, it is natural to require that both mixtures f and $g(\cdot; \boldsymbol{\mu})$ have the most similar clustering information. This information is measured by the probabilities of classification (see (3)). We denote now p_f as the *probability distribution function of the probabilities of classification under mixture* f and $\mathrm{p}_g(\cdot; \boldsymbol{\mu})$ as the *probability distribution function of the probabilities of classification under mixture* $g(\cdot; \boldsymbol{\mu})$. In this manner, a quite natural way for measuring the difference between both f and $g(\cdot; \boldsymbol{\mu})$ clustering property could be the following Kullback-Leibler divergence (p_f being the reference measure):

$$\delta_{\mathrm{KL}}(f, g(\cdot; \boldsymbol{\mu})) = \int_{\mathcal{T}} \mathrm{p}_f(\boldsymbol{t}) \ln \frac{\mathrm{p}_f(\boldsymbol{t})}{\mathrm{p}_g(\boldsymbol{t}; \boldsymbol{\mu})} d\boldsymbol{t} \tag{9}$$

where $\mathcal{T} = \{\boldsymbol{t} : \boldsymbol{t} = (t_1, \ldots, t_{K-1}), t_k > 0, \sum_k t_k < 1\}$. Then, the set $\mathcal{G}$ is defined as

$$\mathcal{G} = \{g : g = g(\cdot; \boldsymbol{\mu}), g \in \arg\min \delta_{\mathrm{KL}}(f, g), f \in \mathcal{F}\}. \tag{10}$$

Somewhere, it is mimicking the idea of Scrucca (2010) and Morris et al. (2013) which imposes (as far as possible) the retention of the overlap of mixture distributions before and after the mapping.

More constrains on Gaussians Another natural requirement should be that $\mathrm{p}_g(\cdot;\boldsymbol{\mu})$ and g are linked by a one-to-one mapping, meaning that for one distribution f, there is a unique distribution $g(\cdot;\boldsymbol{\mu})$ which minimizes (9). This target is reached firstly by setting $d_Y = K-1$ and secondly by setting $\boldsymbol{\mu}_K = \mathbf{0}$, $\mu_{kh} = 0$ if $h > k$, and $\mu_{kk} \geq 0$. This last restriction prevents any rotation and/or translation of $\boldsymbol{y}$ from providing the same distribution $\mathrm{p}_g(\cdot;\boldsymbol{\mu})$ but a different distribution $g(\cdot;\boldsymbol{\mu})$.

For technical convenience, we consider now the following one-to-one mapping Λ between $\boldsymbol{t}$ and a classical transformation of the $\boldsymbol{t}$, which we express hereafter as $\tilde{\boldsymbol{t}}$

$$\Lambda : \boldsymbol{t} = (t_1, \ldots, t_{K-1}) \in \mathcal{T} \mapsto \tilde{\boldsymbol{t}} = (\tilde{t}_1, \ldots, \tilde{t}_{K-1}) \in [0,\infty)^{K-1} \text{ with } \tilde{t}_k = \frac{t_k}{1 - \sum_{\ell=1}^{K-1} t_\ell}. \quad (11)$$

It is essential to note that, by considering mixture $g(\cdot;\boldsymbol{\mu})$, there is also a one-to-one mapping Ψ between $\boldsymbol{y}$ and $\tilde{\boldsymbol{t}}$

$$\Psi(\cdot;\boldsymbol{\mu}) : \boldsymbol{y} \mapsto \tilde{\boldsymbol{t}} \text{ with } \Psi(\boldsymbol{y};\boldsymbol{\mu}) = \left(\frac{\pi_1 \phi_{d_Y}(\boldsymbol{y};\boldsymbol{\mu}_1,\boldsymbol{I})}{\pi_K \phi_{d_Y}(\boldsymbol{y};\boldsymbol{\mu}_K,\boldsymbol{I})}, \ldots, \frac{\pi_{K-1} \phi_{d_Y}(\boldsymbol{y};\boldsymbol{\mu}_{K-1},\boldsymbol{I})}{\pi_K \phi_{d_Y}(\boldsymbol{y};\boldsymbol{\mu}_K,\boldsymbol{I})} \right). \quad (12)$$

Moreover, we have

$$\Psi^{-1}(\tilde{\boldsymbol{t}};\boldsymbol{\mu}) = \boldsymbol{M}^{-1} \begin{pmatrix} \ln\left(\tilde{t}_1 \frac{\pi_K}{\pi_1}\right) + \frac{1}{2}||\boldsymbol{\mu}_1||^2 \\ \vdots \\ \ln\left(\tilde{t}_{K-1} \frac{\pi_K}{\pi_{K-1}}\right) + \frac{1}{2}||\boldsymbol{\mu}_{K-1}||^2 \end{pmatrix} \text{ with } \boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \vdots \\ \boldsymbol{\mu}_{K-1}' \end{pmatrix}, \quad (13)$$

where matrices $\boldsymbol{M}$ and $\boldsymbol{M}^{-1}$ are lower triangular.

3.3 Estimating the Gaussian centers

Invoking a log-likelihood From (10), we consider the distribution $g(\cdot;\boldsymbol{\mu}^*)$ where the centers $\boldsymbol{\mu}^*$ are defined by $\boldsymbol{\mu}^* = \arg\min \delta_{\mathrm{KL}}(f, g(\cdot;\boldsymbol{\mu}))$. Noting that $|\mathrm{Jac}_\Lambda(\boldsymbol{t})|^{-1} = \sum_{k=1}^{K-1} t_k$,

$$\boldsymbol{\mu}^* = \arg\min \int_{[0,\infty)^{K-1}} \tilde{\mathrm{p}}_f(\tilde{\boldsymbol{t}}) \ln \frac{\tilde{\mathrm{p}}_f(\tilde{\boldsymbol{t}})}{\tilde{\mathrm{p}}_g(\tilde{\boldsymbol{t}};\boldsymbol{\mu})} \left(\sum_{k=1}^{K-1} t_k \right) d\tilde{\boldsymbol{t}}, \quad (14)$$

where $\tilde{\mathrm{p}}_f(\cdot)$ and $\tilde{\mathrm{p}}_g(\cdot;\boldsymbol{\mu})$ denote the pdf of $\tilde{\boldsymbol{t}}$ by considering distribution f and $g(\cdot;\boldsymbol{\mu})$, respectively. It is possible to explicitly and easily express the previous $\tilde{\mathrm{p}}_g(\cdot;\boldsymbol{\mu})$ distribution by using the change of variables theorem combined with the linear transformation (13), which leads to the following term, obtained by noting that $|\mathrm{Jac}_{\Psi^{-1}(\cdot;\boldsymbol{\mu})}(\tilde{\boldsymbol{t}})|^{-1} = \prod_{k=1}^{K-1}(\mu_{kk}\tilde{t}_k)^{-1}$,

$$\tilde{\mathrm{p}}_g(\tilde{\boldsymbol{t}};\boldsymbol{\mu}) = g(\Psi^{-1}(\tilde{\boldsymbol{t}};\boldsymbol{\mu});\boldsymbol{\mu}) \prod_{k=1}^{K-1}(\mu_{kk}\tilde{t}_k)^{-1}. \tag{15}$$

Unfortunately, the Kullback-Leibler divergence defined in (14) has generally no closed-form. However, it is easy to independently draw a sample of S ratios of conditional probabilities $\tilde{\mathbf{t}} = (\tilde{\boldsymbol{t}}^{(1)}, \ldots, \tilde{\boldsymbol{t}}^{(S)})$ from $\tilde{\mathrm{p}}_f$. This sample can be used to estimate the previous integral such that maximizing the following normalized (observed-data) log-likelihood function

$$L(\boldsymbol{\mu};\tilde{\mathbf{t}}) = \frac{1}{S}\sum_{s=1}^{S} \ln \tilde{\mathrm{p}}_g(\tilde{\boldsymbol{t}}^{(s)};\boldsymbol{\mu}), \tag{16}$$

is equivalent to solving (14) asymptotically on S.

Maximizing the log-likelihood The log-likelihood (16), combined with (15), entails the pdf of a mixture model. Thus, it can be classically broken down into a normalized complete-data log-likelihood L_{comp} and a normalized entropy term E as follows (Hathaway 1986):

$$L(\boldsymbol{\mu};\tilde{\mathbf{t}}) = L_{\mathrm{comp}}(\boldsymbol{\mu};\tilde{\mathbf{t}}) + E(\tilde{\mathbf{t}}), \tag{17}$$

where, noting $\Lambda^{-1}(\tilde{\boldsymbol{t}}) = (\Lambda_1^{-1}(\tilde{\boldsymbol{t}}), \ldots, \Lambda_{K-1}^{-1}(\tilde{\boldsymbol{t}}))$ the inverse function of Λ,

$$L_{\mathrm{comp}}(\boldsymbol{\mu};\tilde{\mathbf{t}}) = c - \sum_{k=1}^{K-1}\ln\mu_{kk} - \frac{1}{2S}\sum_{s=1}^{S}\sum_{k=1}^{K-1}\Lambda_k^{-1}(\tilde{\boldsymbol{t}}^{(s)})||\Psi^{-1}(\tilde{\boldsymbol{t}}^{(s)};\boldsymbol{\mu}) - \boldsymbol{\mu}_k||^2, \tag{18}$$

$$E(\tilde{\mathbf{t}}) = -\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K-1}\Lambda_k^{-1}(\tilde{\boldsymbol{t}}^{(s)})\ln\Lambda_k^{-1}(\tilde{\boldsymbol{t}}^{(s)}), \tag{19}$$

with a constant term $c = \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K-1}\Lambda_k^{-1}(\tilde{\boldsymbol{t}}^{(s)})\ln\pi_k - \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K-1}\ln\Lambda_k^{-1}(\tilde{\boldsymbol{t}}^{(s)})$. Since the normalized entropy does not depend on $\boldsymbol{\mu}$, an estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}^*$ is obtained only via the maximization of the normalized complete-data likelihood. Note that this maximization is straightforward only if $K = 2$. In such case, we have $\hat{\mu}_1 \in \mathbb{R}$ with $\hat{\mu}_1 = \frac{-1+\sqrt{\frac{1}{S}(\sum_{s=1}^{S} t_{s1})(\sum_{s'=1}^{S} t_{s'1}[\ln(\tilde{t}_{s1}\frac{\pi_2}{\pi_1})]^2)}}{\frac{1}{2S}\sum_{s=1}^{S} t_{1s}}$. Thus, if the overlap between the two components increases (*i.e.*, $t_{s1} \to \frac{1}{2}$ which lead that $\tilde{t}_{s1} \to 1$ and $\frac{\pi_2}{\pi_1} \to 1$) then we have $\hat{\mu} \to 0$. Moreover, when the overlap between the two compoennts decreases, $\mu \to 0$. Note that these remarks stay valid for any model used for clustering.. If the number of components is more than two, a standard Quasi-Newton algorithm should be run with different random initializations, in order to avoid possible local optima. In practice, we use $S = 5000$ which allows for a fast estimation of the centers and stability of the results.

Remark It can be noticed that generative topographic mapping (Bishop et al. 1998) (GTM) could have some similarities with our approach since it is also based on a spherical Gaussian mixture model of the data, estimated through an EM algorithm. However, this fitted distribution is a mixture where the position of the centers of the clusters on the latent space (typically two-dimensional) are defined by advance on a regular grid avoiding any clustering interpretation. Thus GTM is essentially a non-linear dimensionality reduction where no particular clustering focus is taken into account.

4 Final visualization as bivariate spherical Gaussians

4.1 From a multivariate to a bivariate Gaussian mixture

Because g is defined on $\mathbb{R}^{K-1}$, it is inconvenient to draw this distribution if $K \geq 4$. Therefore, we apply an LDA to g to represent this distribution on its most discriminative map

(*i.e.,* eigen value decomposition of the covariance matrix computed on the centers $\hat{\boldsymbol{\mu}}$ by considering the mixture proportions $\boldsymbol{\pi}$), leading to the following bivariate spherical Gaussian mixture $\tilde{g}$:

$$\tilde{g}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \pi_k \phi_2(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}}_k, \boldsymbol{I}), \tag{20}$$

where $\tilde{\boldsymbol{y}} \in \mathbb{R}^2$, $\tilde{\boldsymbol{\mu}} = (\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_K)$ and $\tilde{\boldsymbol{\mu}}_k \in \mathbb{R}^2$. The (standard) percentage of inertia of LDA serves to measure the quality of the mapping from g to $\tilde{g}$. In addition, the accuracy of the mapping from the initial mixture f to the final "ready-to-be-drawn" mixture $\tilde{g}$ can be easily compared through the following difference between the normalized entropy of f and the normalized entropy of $\tilde{g}$, namely

$$\delta_{\mathrm{E}}(f, \tilde{g}) = -\frac{1}{\ln K} \sum_{k=1}^{K} \left\{ \int_{\mathcal{X}} t_k(\boldsymbol{x}; f) \ln t_k(\boldsymbol{x}; f) d\boldsymbol{x} - \int_{\mathbb{R}^2} t_k(\tilde{\boldsymbol{y}}; \tilde{g}) \ln t_k(\tilde{\boldsymbol{y}}; \tilde{g}) d\tilde{\boldsymbol{y}} \right\}. \tag{21}$$

Such a quantity can be easily estimated using empirical values. Its meaning is particularly relevant: if $\delta_{\mathrm{E}}(f, \tilde{g})$ is close to zero then the component overlap conveyed by $\tilde{g}$ (over f) is accurate; if it is close to one, then $\tilde{g}$ strongly underestimates the component overlap of f; if it is close to negative one, then $\tilde{g}$ strongly overestimates the component overlap of f. Thus, $\delta_{\mathrm{E}}(f, \tilde{g})$ serves to evaluate the bias of the visualization.

Remark When the initial data set $\mathbf{x}$ is in the continuous space $\mathcal{X} = \mathbb{R}^d$ and also when the initial clustering relies on a Gaussian mixture f whose covariance matrices are identical, then the proposed mapping is strictly equivalent to applying a LDA to the centers of f.

4.2 Proposal for drawing the bivariate Gaussian mixture

The aim is now to draw the pdf $\tilde{g}$ on the most discriminative map in a manner that highlights as much as possible the overlap between components. Indeed, it is primarily

such information that acted as a guideline to transform f into $\tilde{g}$. The proposed graph will display the following elements:

- **Cluster centers**: the locations of $\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_K$ are materialized by vectors.
- **Cluster spread**: the 95% confidence level is displayed by a black border which separates the area outside the confidence level in white from the area inside the confidence level in gray levels (*i.e.,*, the set Ω is in gray where $\Omega = \{\tilde{\boldsymbol{y}} : g(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}}) > u_\alpha\}$ with u_α such that $\int_\Omega g(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}}) d\tilde{\boldsymbol{y}} = 1 - \alpha$; by default plots are made with $\alpha = 0.05$).
- **Cluster overlap**: curves of iso-probability of the MAP classification are also displayed for different levels ℓ (associated with different gray levels), a curve being composed of the set of $\tilde{\boldsymbol{y}}$ such that

$$\max_{k=1,\ldots,K} \frac{\pi_k \phi_2(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}}_k)}{\tilde{g}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{\mu}})} = \ell. \tag{22}$$

- **Mapping accuracy**: the accuracy of this representation is given by the difference between entropies $\delta_{\mathrm{E}}(f, \tilde{g})$ and also by the percentage of inertia by axis.

4.3 Tutorial on the bivariate spherical Gaussians visualization

We offer here a tutorial for avoiding any misinterpretation of the proposed bivariate spherical Gaussians visualization. It illustrates also its potential great interest for having a fast, easy, unifying and faithful overview of the potentially complex underlying clustering structure. The selected illustrative mixture corresponds to four bivariate Gaussians with non-spherical covariance matrices and different mixing proportions. In this simple and well-known scenario, many standard bivariate illustrations of Gaussians and/or a related data set already exist, of which users are familiar with them. By this way, users would easily understand how to properly analyze the new drawing we propose.

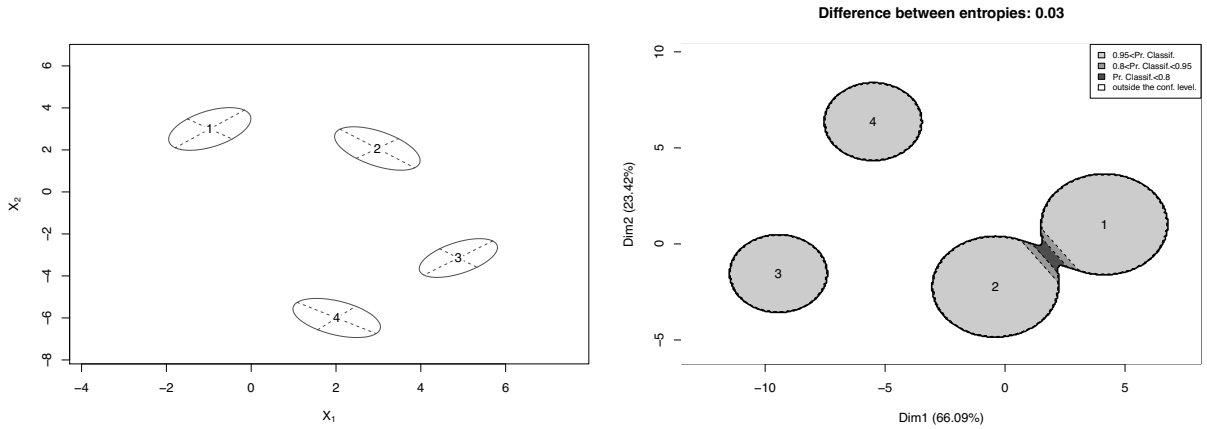
The 1st bivariate ($d_X = 2$) Gaussian layout (called hereafter "1st scenario") is composed of four components ($K = 4$) with mixing proportions $\pi_1 = \pi_2 = 0.4$ and $\pi_3 = \pi_4 = 0.1$, with means $\nu_1 = (-1, 3)$, $\nu_2 = (3, 2)$, $\nu_3 = (5, -3)$, $\nu_4 = (2, -6)$ and with heteroscedastic covariance matrices $\Sigma_1 = \Sigma_3 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and $\Sigma_2 = \Sigma_4 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$. Figure 2a displays isodensity curves of the related mixture provided by the classical R package `mclust` (Scrucca et al. 2016). Just the component number has been manually overlayed on the means. Many other packages are expected to offer similar visualization choices. Figure 2b displays the proposed bivariate spherical Gaussian visualization associated to this 1st mixture scenario. Note that this Gaussian representation is really spherical, even if it can appear distorted due to the axes scaling. First of all the difference between the entropies has to be checked. Its low absolute value (0.03) indicates that the cluster overlap displayed on the figure is globally accurate. Thus the following comments on the initial heteroscedastic mixture we will make through this new spherical representation are valid:

- Axes meaning: the 1st axis is the most discriminative one provided by the LDA mapping (66.09% of the discriminant power). The first two axes sum to $66.09 + 23.41 = 89.50\%$ of the discriminant power, thus most of the discriminant information is present on this two dimensional mapping.

- Mixing proportions: the confidence areas in gray color are directly related to the mixing proportions. Thus it appears immediately that components 1 and 2 are more populous than the two others.

- Cluster overlap: it clearly appears also that components 1 and 2 overlap much more than components 3 and 4 do. This fact does not appear clearly at all on Figure 2a since mixing proportions are not involved in the iso-density representation. More

generally, separation of all couples of components appears to be faithful. For instance, components 2 and 4 (and also components 1 and 3) are the most separated ones.

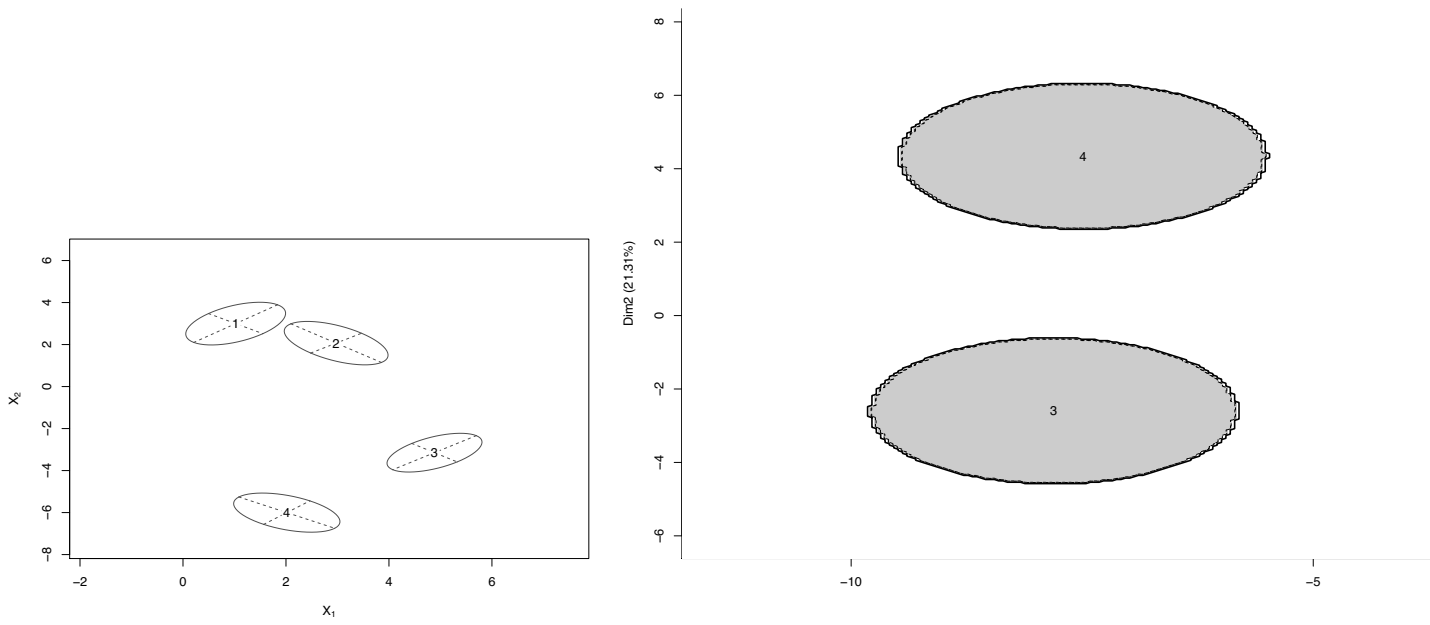


(a) classical isodensity curves representation (b) proposed spherical-like representation

Figure 2: Representation of the clusters for the first scenario.

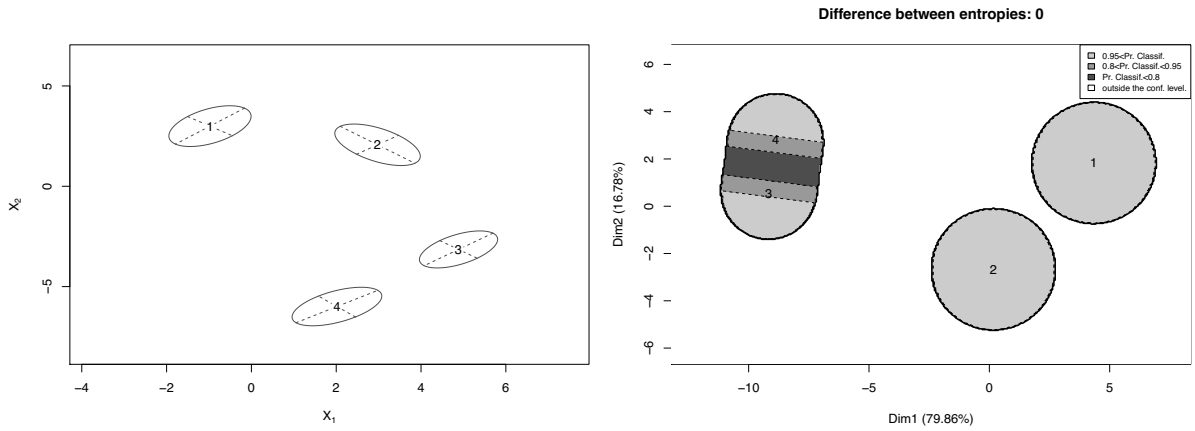
Figure 3 considers a similar case to 1st scenario but where components 1 and 2 are closer with regards to their means, so their overlap has increased. Indeed, here $\nu_1 = (1, 3)$, Figure 3b has now a lower displaying accuracy compared to Figure 2b since difference between the entropies is 0.15. However its absolute value is sufficiently close to zero and far from one (its maximum theoretical value) to allow faithful interpretation of the overall components displaying. Figure 3b clearly indicates that components 1 and 2 overlap significantly more, what is really the fact in the underlying experimental design.

Figure 4 considers a similar case to 1st scenario but where components 3 and 4 are closer with regards to their covariance matrices, so their overlap has increased. Indeed, here $\Sigma_4 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. This 3nd scenario is particularly interesting since the spherical



(a) classical isodensity curves representation (b) proposed spherical-like representation

Figure 3: Representation of the clusters for the second scenario.



(a) classical isodensity curves representation (b) proposed spherical-like representation

Figure 4: Representation of the clusters for the third scenario.

representation is unable to distort its covariance matrices (they are fixed to be spherical and identical). Consequently, only means of these spherical Gaussians can be distorted to faithfully represent the corresponding new overlap. Figure 4b shows that this means adaptation was successfully enough since difference between the entropies is very close to zero. And it can be seen on the same figure that components 3 and 4 overlap very significantly, as expected.

Finally, Figure 6 considers the same scenario that Figure 4 where the components proportions are equal. It illustrates that the size of the gray areas around the centers reflects the size of the components.

4.4 Continuation of the running example

We now illustrate the previous visualization proposition on the running example. Figure 5 is the component interpretation graph obtained for the congressional voting records. It

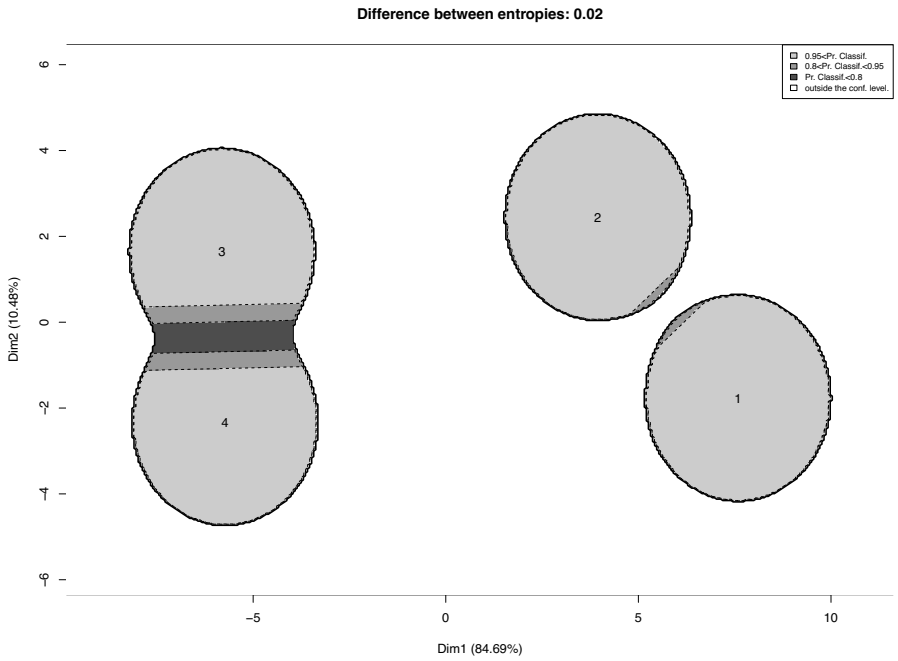


Figure 5: proposed spherical-like representation for the fourth scenario.

presents the Gaussian-like component overlap on the most discriminative map. In this way, it provides in a more concise way than a traditional confusion table the overlap information of the initial mixture f. Note that the mapping of f on this graph is accurate because the difference between entropies is almost zero (*i.e.,* $\delta_{\mathrm{E}}(f, \tilde{g}) = 0.01$). For instance, this figure also shows that the components with most observations (*i.e.,* components three and four) are composed of strongly different Congressmen. Indeed, the overlap between these components is almost zero. Moreover, component one contains Congressmen which are more moderated that Congressmen of components three and four.

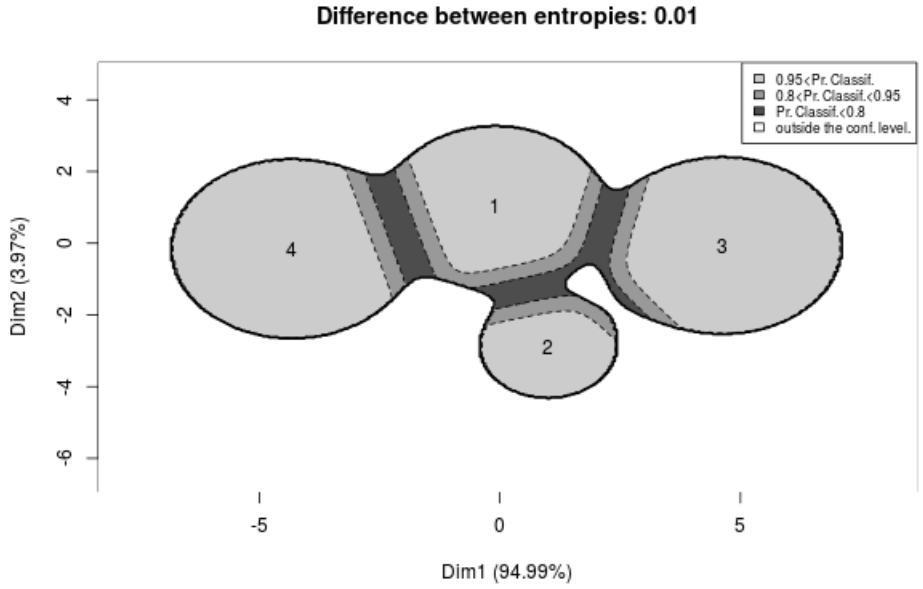


Figure 6: Component interpretation graph of the congressional voting records.

5 Proposal for drawing a *pseudo* bivariate scatter plot

5.1 From pdf visualization to individual visualization

We have limited our attention to the mapping of the initial cluster pdf f, described by (5), intentionally discarding the mapping of the initial individual data set $\mathbf{x}$, described by (4). We have already discussed that the pdf mapping (5) can be a by-product of the individual mapping (4). However, the reverse is mathematically impossible, the distributional information being weaker than the random variable information.

Nevertheless, a *pseudo* scatter plot $\mathbf{y}$ of $\mathbf{x}$ can be mapped onto $\mathbb{R}^{K-1}$ by transforming the ratios of probabilities $\Lambda(\boldsymbol{t}_i(f))$, associated with $\boldsymbol{x}_i$ by f, into values $\boldsymbol{y}_i$ through the reverse application of $\Psi^{-1}(\cdot;\hat{\boldsymbol{\mu}})$ associated with $g(\cdot;\hat{\boldsymbol{\mu}})$, namely $\boldsymbol{y}_i = \Psi^{-1}(\Lambda(\boldsymbol{t}_i(f));\hat{\boldsymbol{\mu}})$ $(i = 1,\ldots,n)$. Then, each observation $\boldsymbol{y}_i$ is projected on the LDA map, leading to a *pseudo* scatter plot $\tilde{\mathbf{y}} = (\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_n)$, with each $\tilde{\boldsymbol{y}}_i \in \mathbb{R}^2$.

We use the term "pseudo" for $\tilde{\mathbf{y}}$ (or for $\mathbf{y}$) because some caution has to taken in order to avoid misunderstanding. Indeed, the distribution of $\tilde{\mathbf{y}}$ is expected to be different from $g(\cdot;\tilde{\boldsymbol{\mu}})$, the essential property of $\tilde{\mathbf{y}}$ being to respect as far as possible the conditional probabilities $\mathbf{t}(f)$ associated to $\mathbf{x}$, not to respect as far as possible the distribution f itself. In fact, only when f corresponds to a spherical Gaussian mixture do distributions of $\tilde{\mathbf{y}}$ and of $g(\cdot;\tilde{\boldsymbol{\mu}})$ match.

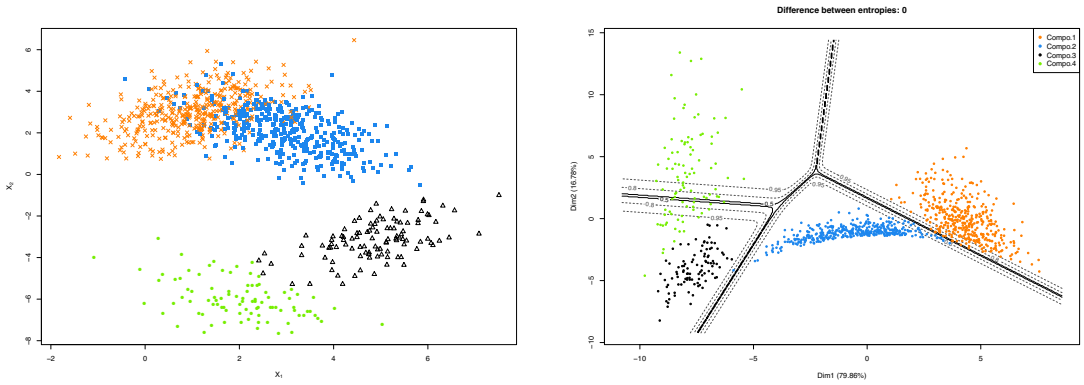
Such remarks strongly affect the related drawing we propose for the scatter plot $\tilde{\mathbf{y}}$:

- **Data drawing**: display $\tilde{\mathbf{y}}$ on the LDA best discriminative map as dots of different colors representing the partition membership $\mathbf{z}$.
- **Conditional probabilities**: information about the uncertainty of classification is given by the curves of iso-probability of classification.

- **Mapping accuracy**: again, the accuracy of this representation is given by the difference between entropies $\delta_{\mathrm{E}}(f, \tilde{g})$ and also by the percentage of inertia by axis.
- **No pdf overlay**: do not display $\tilde{\mathbf{y}}$ simultaneously with $\tilde{g}(.; \tilde{\boldsymbol{\mu}})$ to avoid misunderstanding; therefore use another graph.

5.2 Tutorial on the pseudo bivariate scatter plot visualization

Figure 6a displays, in a classical way, a sample of size n=1000 from the 3rd scenario described in Section 4.3. Figure 6b displays the related pseudo scatter plot we propose. The LDA map is exactly the same between this figure and Figure 4b. However, some comments are required for avoiding misinterpretation of this new plot.



(a) classical bivariate scatter plot (b) proposed bivariate pseudo scatter plot

Figure 7: Scatter plot related to the 3nd scenario.

Here the scatter plot is not necessarily Gaussian (spherical or other), phenomenon that appears clearly. Indeed, remember that the only property of the initial mixture which is preserved through the procedure we propose is only the conditional membership distribu-

tion (or in short its "overlapping") under the constraint that this conditional distribution is a by-product of a spherical Gaussian mixture. Thus, each data sample drawn on Figure 6b has to be seen as a faithful representation of its conditional membership representation under the spherical constraint, but absolutely not a faithful representation of its mixture distribution. The interest is to quicky access to the membership uncertainty of each individual, what becomes also clearer by the borderlines displayed on the figure. Notice obviously that this membership interpretation is accurate as soon as the difference between the entropies is not far from zero (in absolute value), what is the case for this particular scenario.

5.3 Continuation of the running example

Figure 7 displays the scatter plot of the observation memberships obtained on the congressional voting records. It overlays on the most discriminative map the curve of iso-probabilities of classification and the cloud of observations. Three levels of probabilities of classification are considered (0.95, 0.80 and 0.50) and observations are represented with the label of the component maximizing the posterior probability of classification. This plot serves to focus on specific observations and, for instance, to detect observations classified with a high uncertainty of classification. Note that some points which are classified in component two (the blue points) are located in the area containing the observations of component three. This is a standard phenomenon in LDA. Indeed, in such a case, if the maximum *a posteriori* rule is applied on the native space (*i.e.,* $\mathbb{R}^{K-1}$) than its results can be different to the results of maximum *a posteriori* rule when applied on the low-dimension space (*i.e.,* $\mathbb{R}^2$).

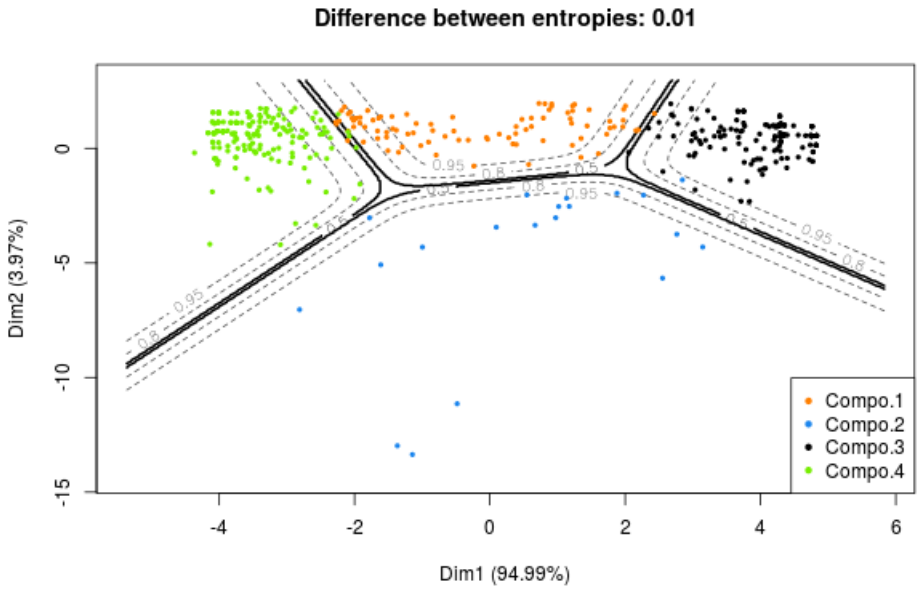


Figure 8: Scatter plot of the observation memberships of the congressional voting records.

6 Numerical illustrations for complex data

We present applications of the visualization method on three real data sets composed of complex features (mixed, functional and network data). They illustrate the ability of the method to deal with extremely different kinds of data and of mixtures, without any new specific development. Results are obtained by the R package ClusVis which implements the visualization method.

6.1 Mixed data: Contraceptive method choice

Data This dataset **x** is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey (Lim et al. 2000). It described 1473 Indian women with two numerical variables

(age and number of children) and eight categorical variables (education level, education level of the husband, religion, occupation, occupation of the husband, standard-of-living index and media exposure).

Model used to cluster These mixed data are clustered by a mixture f assuming that variables are independent within components (Moustaki & Papageorgiou 2005). Within a component, the continuous variables follow Gaussian distributions and categorical variables follow multinomial distributions. Maximum likelihood inference is performed by the R package Rmixmod (Lebret et al. 2015). Model selection is done by the BIC criterion which detects six components.

Model drawing Figure 8 presents the component interpretation graph obtained for the contraceptive method choice data. It shows overlaps between component one, two and three. Moreover, components four and five are significantly different from component six. Such a visualization is in accordance with a fine study of Table 1, which presents the parameters of the continuous variables. Indeed, we can see that components one, two and three are all composed of middle-age women who have many children. On the contrary, components four and five are composed of young women who have few children. Finally, component six is composed of the oldest women. Therefore, the first axis can be interpreted as the age of the women (left side is composed of older women than on the right side). Finally, the second axis distinguishes components two, four and six from the others. As shown in Table 2, these three components have the same mode for the eight categorical variables.

Scatter plot drawing The scatter plot of the observation memberships is presented in Figure 9. The overlap between components one and three is obvious. Note that, on this

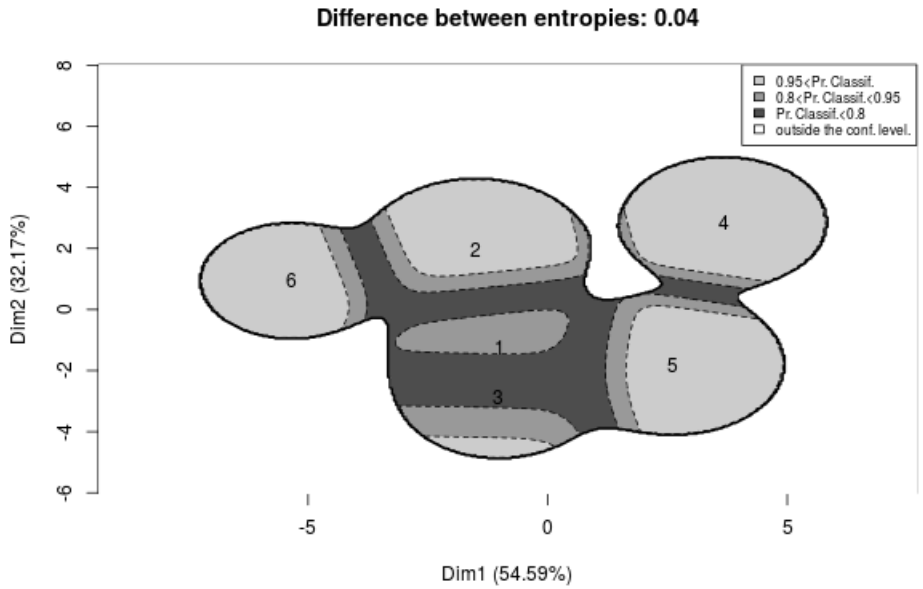


Figure 9: Component interpretation graph of the Contraceptive method choice.

	Age		Number of children	
	Mean	Variance	Mean	Variance
Component 1	35	30	4	4
Component 2	35	22	3	2
Component 3	40	42	5	9
Component 4	25	10	1	1
Component 5	24	13	2	1
Component 6	45	7	5	8

Table 1: Parameters of the continuous variables for the Contraceptive method choice.

	education level	husband's education level	religion	occupation	husband's occupation	standard-of-living index	media exposure
Component 1	3	3	2	2	3	4	1
Component 2	4	4	2	2	1	4	1
Component 3	1	2	2	2	3	3	1
Component 4	4	4	2	2	1	4	1
Component 5	3	3	2	2	3	3	1
Component 6	4	4	2	2	1	4	1

Table 2: Modes of the categorical variables for the Contraceptive method choice.

figure, some observations classified under component one are projected on a location where the MAP rule would classify them under component three. However, on the space $\mathbb{R}^5$, the probabilities of classification are respected precisely. But this well-known phenomenon is due to the projection of the observations $\boldsymbol{y}_i$ from $\mathbb{R}^5$ to $\mathbb{R}^2$ when projecting a discriminative rule.

6.2 Functional data: Bike sharing system

Data We consider now the study of the Bike sharing system data presented by Bouveyron et al. (2015). We analyze station occupancy data collected over the course of one month on the bike sharing system in Paris. The data were collected over 5 weeks, between February, 24 and March, 30, 2014, at 1 189 bike stations. The station status information, in terms of available bikes and docks, were downloaded every hour during the study period for the seven systems from the open-data APIs provided by the JCDecaux company. To accommodate the varying stations sizes (in terms of the number of docking points), Bouveyron et al. (2015) normalized the number of available bikes by station size and obtained a loading profile for each station. The final data set contains 1 189 loading profiles, one per station,

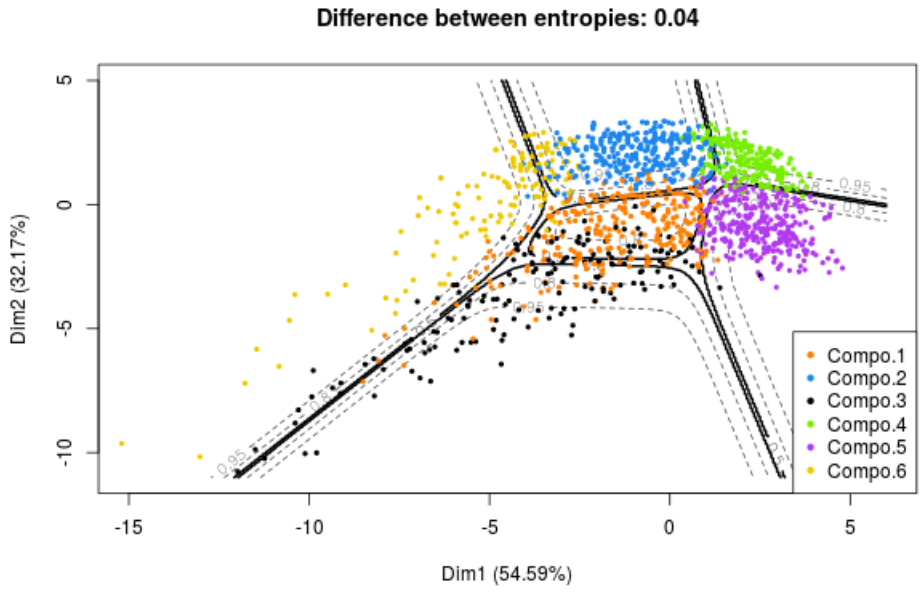


Figure 10: Scatter plot of the observation memberships of the Contraceptive method choice.

sampled at 1 448 time points. Note that the sampling is not perfectly regular; there is one hour, on average, between the two sample points. The daily and weekly habits of inhabitants introduce a periodic behavior in the BSS station loading profiles, with a natural period of one week. It is thus natural to use a Fourier basis to smooth the curves, with basis functions corresponding to sine and cosine functions of periods equal to fractions of this natural period of the data. Using such a procedure, the profiles of the stations were projected on the basis of 25 Fourier functions.

Model used to cluster We conduct a model-based clustering of these functional data (Bouveyron et al. 2015) using the R package FunFEM (Bouveyron 2015) . The parameters of the model presented by Bouveyron et al. (2015) (*i.e.*, $K = 10$ and $DFM_{[\alpha_{kj}\beta]}$ model) are

estimated. Figure 10 presents the curves for the 10 components based on the MAP rule.

Model drawing Figure 11 presents the component interpretation graph obtained for the bike sharing system data. The representation has good accuracy, because the difference between entropies is small (*i.e.,* $\delta_{\mathrm{E}}(f, \tilde{g}) = -0.03$). It shows a strong similarity between components three and four. In Figure 10, we can see that the curves classified in these components are similar (high values with the same phase). Component two and six overlap because they have a very low amplitude. Moreover, Figure 11 shows that component seven is the most isolated one. This component corresponds to the group that Bouveyron et al. (2015) called *empty stations*. Finally, components eight and nine are significantly different because they have a phase opposition. Indeed, these components are at opposite locations on this figure. The same remark applies for components one and eight as well. In fact, the reader can easily "plays" with Figure 10 and Figure 11 for checking similarities and differences between all components.

Scatter plot drawing The scatter plot of the observation memberships is presented in Figure 12. It confirms the interpretation of Figure 11. Indeed, the observations classified in components three and four are well-mixed. Similarily, one can observe an overlap between components two and six. Finally, the observations classified in component seven are isolated.

6.3 Network data: French political blogosphere

Data We consider the clustering of the French political blogosphere network (Zanghi et al. 2008). Data consist of a single day snapshot of over 1 100 political blogs automatically extracted on October, 14th, 2006 and manually classified by the "Observatoire Presidentielle"

project. This project is the result of collaborative work by RTGI SAS and Exalead and aims at analyzing the French presidential campaign on the web. In this data set, nodes represent hostnames (a hostname contains a set of pages) and edges represent hyperlinks between different hostnames. If several links exist between two different hostnames, Zanghi et al. (2008) subsum them into a single one. Note that intra-domain links can be considered if hostnames are not identical. Finally, in this experimentation we consider that edges are not directed which is not realistic but which does not affect the interpretation of the groups. This network presents an interesting community-based organization due to the existence of several political parties and commentators. We assume that authors of these blogs tend to link, blogs with similar political positions as a result of their political affinities.

Model used to cluster We use the graph clustering via Erdös-Rényi mixture proposed by Zanghi et al. (2008) and implemented on the R package MIXER. As proposed by these authors, we consider $K = 6$ components. The confusion matrix between the component memberships and the political party memberships is given in Table 3.

Model drawing Figure 13 presents the component interpretation graph obtained for the French political blogosphere data. The graph slightly over-represents the component overlaps (*i.e.,* $\delta_{\text{E}}(f, \tilde{g}) = -0.216$). Indeed, the (normalized) entropy of f is equal to 0.016 while the entropy of the projection of g into the most discriminative space is equal to 0.221. Note that this difference between entropies is due to the projection of the data from $\mathbb{R}^5$ to $\mathbb{R}^2$. Indeed, the entropy of g (in $\mathbb{R}^5$) is closed to those of f with a value of 0.004. The loss of information due to the data projection can also be detected by the inertia, because only 56.76% of the inertia is represented by this most discriminative map. Therefore, the components overlaps should be interpreted with more caution than in the

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Cap21	2	0	0	0	0	0
Commentateurs Analystes	10	0	0	1	0	0
FN - MNR - MPF	2	0	0	0	0	0
Les Verts	7	0	0	0	0	0
PCF - LCR	7	0	0	0	0	0
PS	31	0	0	0	26	0
Parti Radical de Gauche	11	0	0	0	0	0
UDF	1	1	0	30	0	0
UMP	2	25	11	2	0	0
liberaux	0	1	0	0	0	24

Table 3: Confusion matrix between the component memberships and the political party memberships.

previous examples, where the differences between entropies were close to zero.

The graph shows that components three and six overlap significantly. This result is natural because component three mainly comprises UMP members ("French Republican") and component six is composed of supporters of economic-liberalism. Finally, component one, which comprises politicians from different political parties, is the most isolated.

Scatter plot drawing Figure 14 presents the scatter plot of the observation memberships. It confirms the proximity between components three and six.

7 Conclusion

We presented a generic method for visualizing the results of a model-based clustering in a "Gaussian way". This method allows for visualization of any model-based clustering made on any type of data, because it is only based on the distribution of classification probabilities. It permits to interpret the results of a model-based clustering but not to select the best clustering method (choosing a clustering method has to be performed before through a classical model selection process). In this way, it is not an exploratory visualization method, as such methods are often dedicated to.

This method produces two graphs. The first graph allows for the component interpretation through all component overlaps. The second graph represents a scatter plot of the observations and many curves of iso-probabilities of classification. It serves to focus on the classification of specific observations and to quantify the risk of misclassification. Finally, the accuracy of the procedure can be measured by taking the difference between the normalized entropies obtained by the model used to cluster and by the model defined on the visualization map.

The proposed procedure has been developed by considering that the model used to visualize is a constrained Gaussian mixture. Obviously, other continuous distributions could be considered. However, these distributions must define a one-to-one relation between the space of the probability of classification and the continuous space. If several distributions compete, then the best distribution could be the distribution that leads to minimization of the Kullback-Leibler divergence $\delta_{\mathrm{KL}}(f, g)$. Alternatively, because there is a step of LDA-like projection, the best distribution could be the distribution that minimizes the difference between the normalized entropies obtained by f and by the projected distribution $\tilde{g}$, namely $\delta_{\mathrm{E}}(f, \tilde{g})$. Finally, if non-Gaussian mixtures are considered, it is crucial that the resulting

graph presenting the component overlaps is still meaningful and does not entail excessively boring calculus. In particular, it could be meaningful to explore non-unimodal component candidates.

References

Ambroise, C. & Matias, C. (2012), 'New consistent and asymptotically normal parameter estimates for random-graph mixture models', *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74**(1), 3–35.
URL: *http://dx.doi.org/10.1111/j.1467-9868.2011.01009.x*

Audigier, V., Husson, F. & Josse, J. (2016*a*), 'Multiple imputation for continuous variables using a bayesian principal component analysis', *Journal of Statistical Computation and Simulation* **86**(11), 2140–2156.

Audigier, V., Husson, F. & Josse, J. (2016*b*), 'A principal component method to impute missing values for mixed data', *Advances in Data Analysis and Classification* **10**(1), 5–26.

Banfield, J. & Raftery, A. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**(3), 803–821.
URL: *http://dx.doi.org/10.2307/2532201*

Benaglia, T., Chauveau, D. & Hunter, D. R. (2009), 'An em-like algorithm for semi- and nonparametric estimation in multivariate mixtures', *Journal of Computational and Graphical Statistics* **18**, 505—526.

Biernacki, C. (2017), Mixture models, *in* J.-J. Droesbeke, G. Saporta & C. Thomas-Agnan,

eds, 'Choix de modèles et agrégation', Technip.
URL: *https://hal.inria.fr/hal-01252671*

Biernacki, C., Celeux, G. & Govaert, G. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(7), 719–725.

Bishop, C. M., Svensén, M. & Williams, C. K. (1998), 'Gtm: The generative topographic mapping', *Neural computation* **10**(1), 215–234.

Bouveyron, C. (2015), *funFEM: Clustering in the Discriminative Functional Subspace*. R package version 1.1.
URL: *https://CRAN.R-project.org/package=funFEM*

Bouveyron, C., Côme, E. & Jacques, J. (2015), 'The discriminative functional mixture model for a comparative analysis of bike sharing systems', *Ann. Appl. Stat.* **9**(4), 1726–1760.
URL: *http://dx.doi.org/10.1214/15-AOAS861*

Bouveyron, C. & Jacques, J. (2011), 'Model-based clustering of time series in group-specific functional subspaces', *Advances in Data Analysis and Classification* **5**(4), 281–300.

Celeux, G. & Govaert, G. (1991), 'Clustering criteria for discrete data and latent class models', *Journal of Classification* **8**(2), 157–176.
URL: *https://doi.org/10.1007/BF02616237*

Celeux, G. & Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern recognition* **28**(5), 781–793.

Chavent, M., Kuentz-Simonet, V. & Saracco, J. (2012), 'Orthogonal rotation in pcamix', *Advances in Data Analysis and Classification* **6**(2), 131–146.

Chen, K. & Lei, J. (2015), 'Localized functional principal component analysis', *J. Amer. Statist. Assoc.* **110**(511), 1266–1275.
URL: *http://dx.doi.org/10.1080/01621459.2015.1016225*

Cox, T. & Cox, M. (2001), *Multidimensional Scaling*, Chapman and Hall.

Daudin, J.-J., Picard, F. & Robin, S. (2008), 'A mixture model for random graphs', *Stat. Comput.* **18**(2), 173–183.
URL: *http://dx.doi.org/10.1007/s11222-007-9046-7*

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of eugenics* **7**(2), 179–188.

Gollini, I. & Murphy, T. (2014), 'Mixture of latent trait analyzers for model-based clustering of categorical data', *Statistics and Computing* **24**(4), 569–588.

Goodman, L. (1974), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**(2), 215–231.

Greenacre, M. (2017), *Correspondence analysis in practice*, CRC press.

Hathaway, R. J. (1986), 'Another interpretation of the em algorithm for mixture distributions', *Statistics and Probability Letters* **4**, 53–56.

Hennig, C. (2004), 'Asymmetric linear dimension reduction for classification', *Journal of Computational and Graphical Statistics* **13**(4), 930–945.

Jacques, J. & Preda, C. (2014), 'Model-based clustering for multivariate functional data', *Comput. Statist. Data Anal.* **71**, 92–106.
URL: *http://dx.doi.org/10.1016/j.csda.2012.12.004*

Jajuga, K., Sokołowski, A. & Bock, H. (2002), *Classification, clustering and data analysis: recent advances and applications*, Springer Verlag.

Josse, J., Chavent, M., Liquet, B. & Husson, F. (2012), 'Handling missing values with regularized iterative multiple correspondence analysis', *Journal of classification* **29**(1), 91–116.

Josse, J., Pagès, J. & Husson, F. (2011), 'Multiple imputation in principal component analysis', *Advances in data analysis and classification* **5**(3), 231–246.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological cybernetics* **43**(1), 59–69.

Kosmidis, I. & Karlis, D. (2015), 'Model-based clustering using copulas with applications', *Statistics and Computing* pp. 1–21.
URL: *http://dx.doi.org/10.1007/s11222-015-9590-5*

Larose, C. (2015), Model-Based Clustering of Incomplete Data, PhD thesis, University of Connecticut.

Lê, S., Josse, J., Husson, F. et al. (2008), 'Factominer: an r package for multivariate analysis', *Journal of statistical software* **25**(1), 1–18.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. & Govaert, G. (2015), 'Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library', *Journal of Statistical Software* **67**(6), 241–270.

Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (2000), 'A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms', *Machine learning* **40**(3), 203–228.

Marbac, M., Biernacki, C. & Vandewalle, V. (2016), 'Latent class model with conditional dependency per modes to cluster categorical data', *Advances in Data Analysis and Classification* **10**(2), 183–207.

Marbac, M., Biernacki, C. & Vandewalle, V. (2017), 'Model-based clustering of gaussian copulas for mixed data', *Communications in Statistics - Theory and Methods* **46**(23), 11635–11656.

Mazo, G. (2017), 'A semiparametric and location-shift copula-based mixture model', *Journal of Classification* **34**(3), 444–464.

McLachlan, G. & Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.

McNicholas, P. (2016), *Mixture model-based classification*, CRC Press.

McNicholas, P. & Murphy, T. (2008), 'Parsimonious Gaussian mixture models', *Stat. Comput.* **18**(3), 285–296.
URL: *http://dx.doi.org/10.1007/s11222-008-9056-0*

McParland, D. & Gormley, I. C. (2016), 'Model based clustering for mixed data: clustmd', *Advances in Data Analysis and Classification* **10**(2), 155–169.

Morris, K., McNicholas, P. & Scrucca, L. (2013), 'Dimension reduction for model-based clustering via mixtures of multivariate t-distributions', **7**, 321–338.

Moustaki, I. & Papageorgiou, I. (2005), 'Latent class models for mixed variables with applications in archaeometry', *Computational statistics & data analysis* **48**(3), 659–675.

Punzo, A. & Ingrassia, S. (2016), 'Clustering bivariate mixed-type data via the cluster-weighted model', *Computational Statistics* **31**(3), 989–1013.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, Springer Series in Statistics, second edn, Springer, New York.

Samé, A., Chamroukhi, F., Govert, G. & Aknin, P. (2011), 'Model-based clustering and segmentation of time series with changes in regime', *Advances in Data Analysis Classification* **5**, 301–321.

Schlimmer, J. (1987), Concept acquisition through representational adjustment, PhD thesis, Department of Information and Computer Science, University of California.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Scrucca, L. (2010), 'Dimension reduction for model-based clustering', *Statistics and Computing* **20**(4), 471–484.
URL: *http://dx.doi.org/10.1007/s11222-009-9138-7*

Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. (2016), 'mclust 5: clustering, classification and density estimation using Gaussian finite mixture models', *The R Journal* **8**(1), 205–233.
URL: *https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf*

Van der Heijden, P. & Escofier, B. (2003), 'Multiple correspondence analysis with missing data', *Analyse des correspondances. Recherches au cżur de l'analyse des donnees* pp. 152–170.

Verbanck, M., Josse, J. & Husson, F. (2015), 'Regularised pca to denoise and visualise data', *Statistics and Computing* **25**(2), 471–486.

Vesanto, J. & Alhoniemi, E. (2000), 'Clustering of the self-organizing map', *IEEE Transactions on neural networks* **11**(3), 586–600.

Xanthopoulos, P., Pardalos, P. M. & Trafalis, T. B. (2013), 'Linear Discriminant Analysis', *Robust Data Mining* pp. 27–33.

Young, F. W. (1987), *Multidimensional scaling: History, theory, and applications*, Lawrence Erlbaum Associates.

Zanghi, H., Ambroise, C. & Miele, V. (2008), 'Fast online graph clustering via erdős–rényi mixture', *Pattern Recognition* **41**(12), 3592 – 3599.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320308002483*

Zhou, L. & Pan, H. (2014), 'Principal component analysis of two-dimensional functional data', *J. Comput. Graph. Statist.* **23**(3), 779–801.
URL: *http://dx.doi.org/10.1080/10618600.2013.827986*

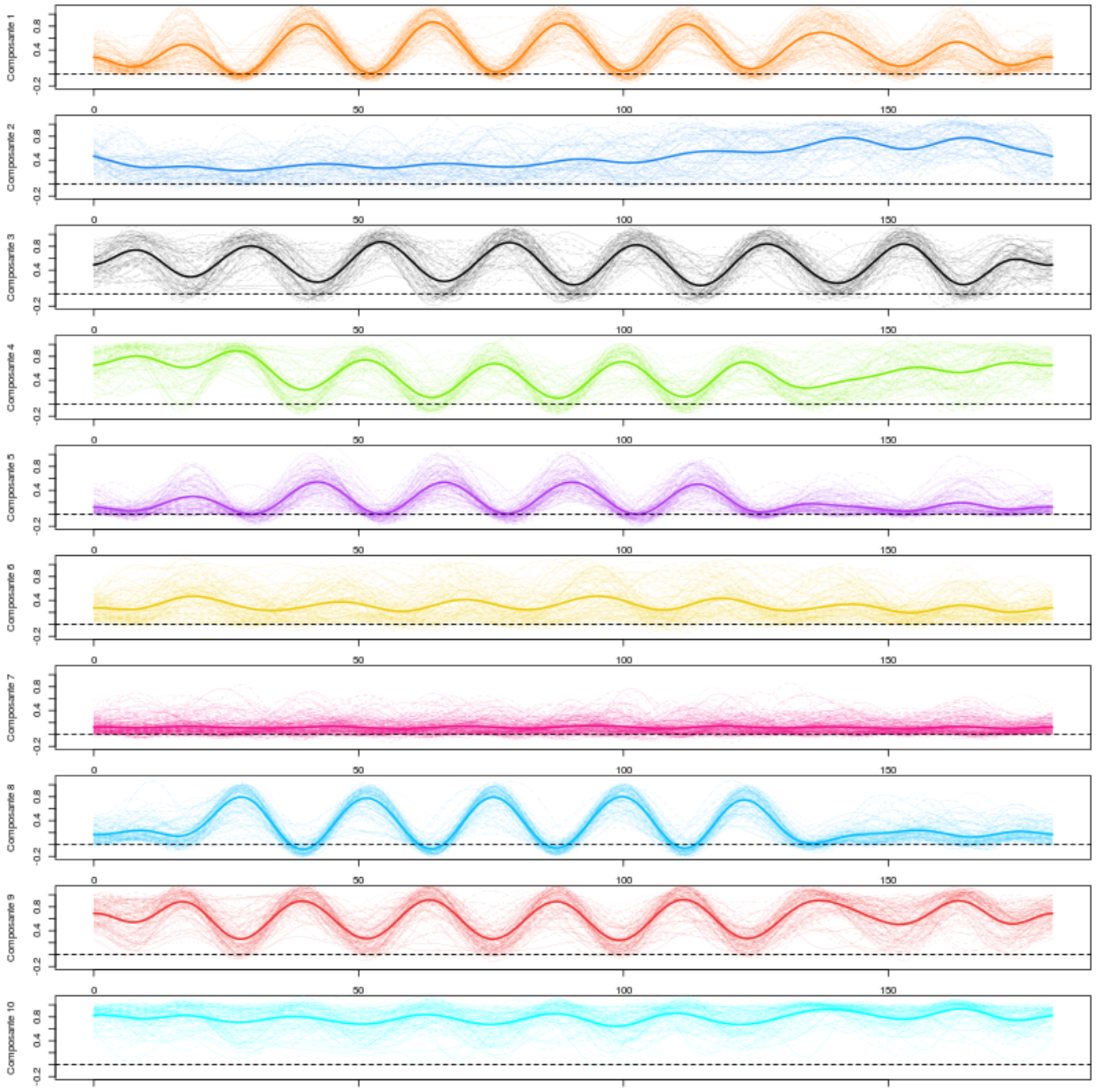


Figure 11: Partition among the 1 189 bike stations in Paris (each row corresponds to a single component and gives the mean curve in bold and observations belonging to this component in thin).

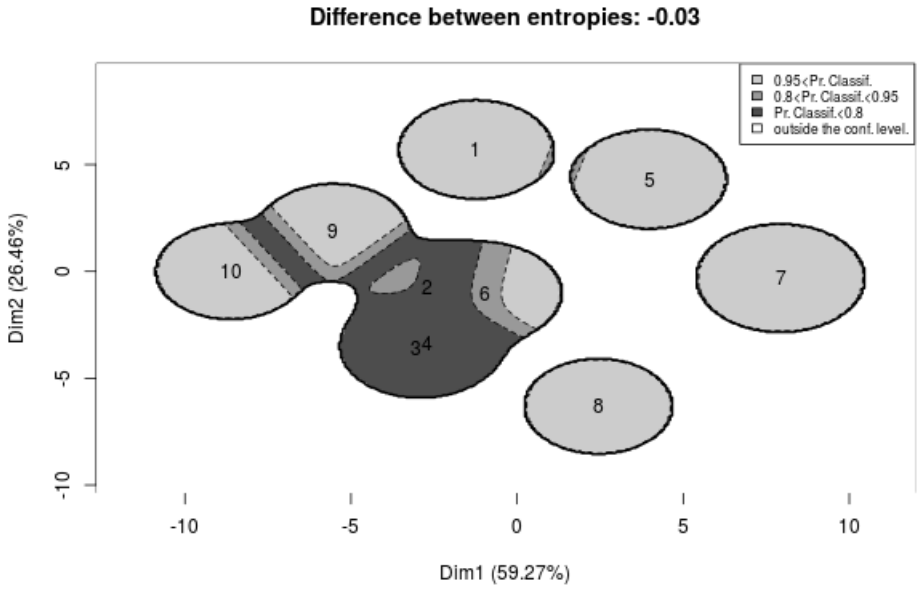


Figure 12: Component interpretation graph of the bike sharing system.

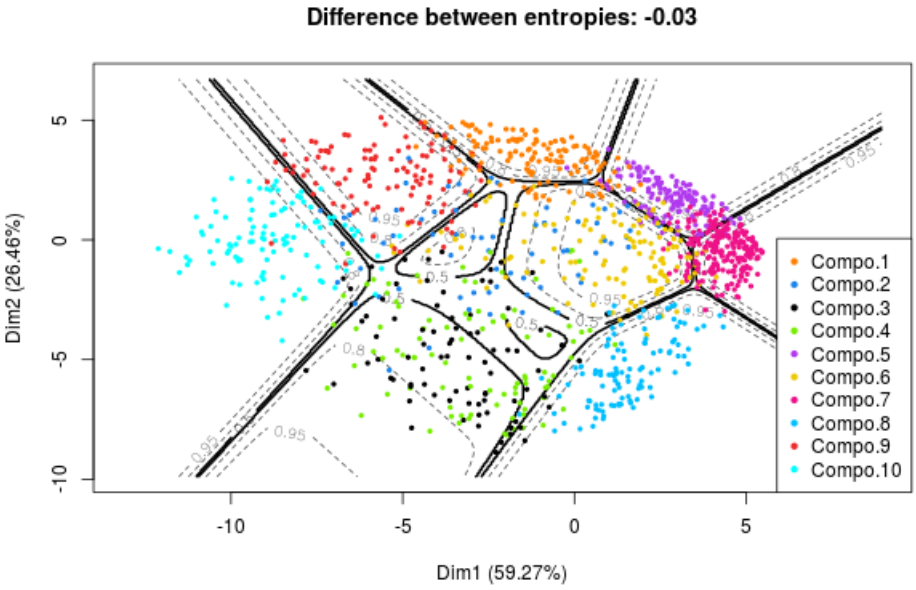


Figure 13: Scatter plot of the observation memberships of the bike sharing system.

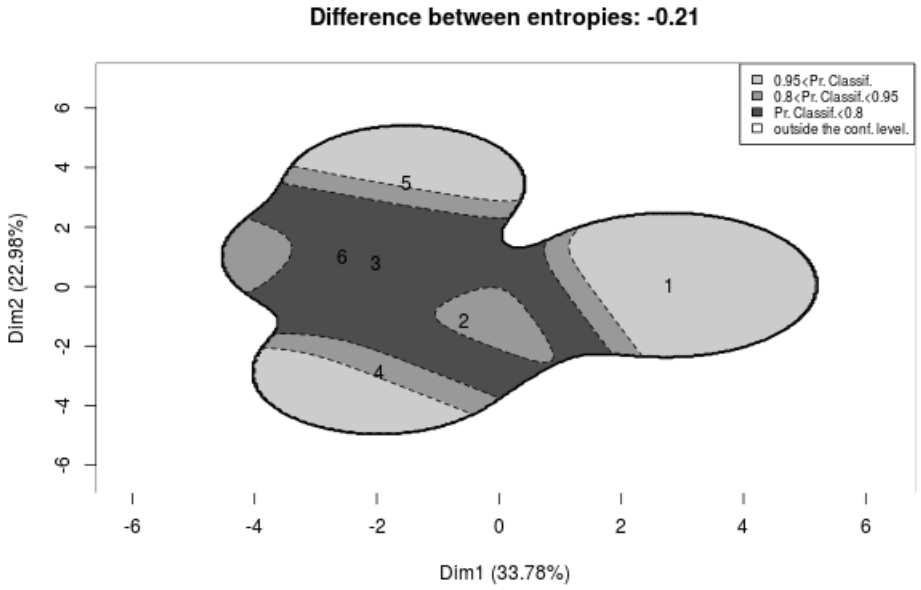


Figure 14: Component interpretation graph of the French political blogosphere.

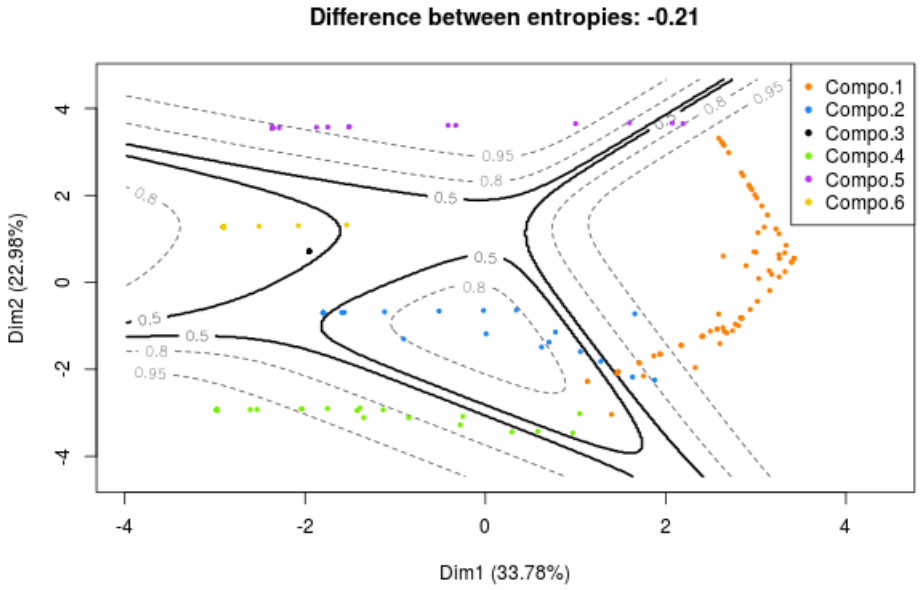


Figure 15: Scatter plot of the observation memberships of the French political blogosphere.