Clustering: from modeling to visualizing Mapping clusters as spherical Gaussians

Christophe Biernacki, Vincent Vandewalle, Matthieu Marbac

Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

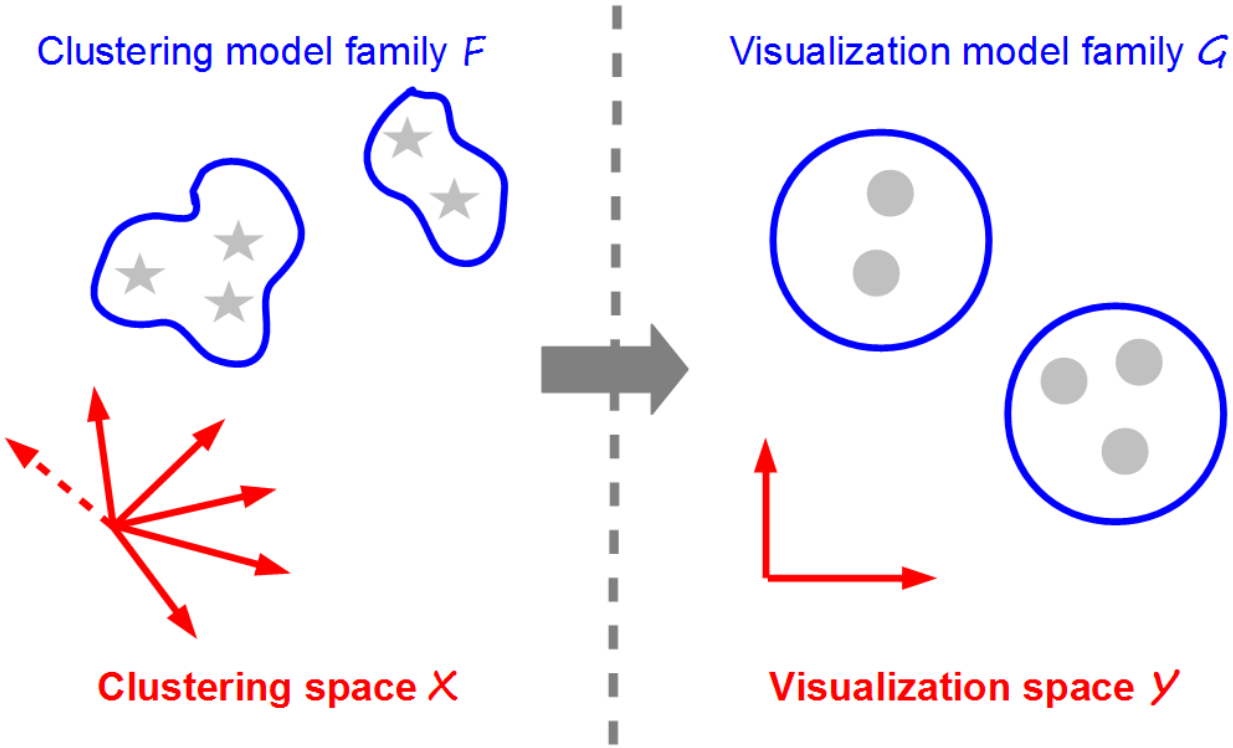
C. Biernacki

with M. Marbac-Lourdelle and V. Vandewalle

SFC'2023
Rencontres de la Société Francophone de Classification
6-7 juillet 2023, Strasbourg, France

Take home message

Traditionally: spaces for visualizing clusters are fixed for their user-convenience
Natural extension: models for visualizing clusters should follow the same principle!



Outline

Model-based clustering: pitch[1]

- Data set: $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, each $\boldsymbol{x}_i \in \mathcal{X}$ with d_X variables
- Partition (unknown): $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ with binary notation $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$
- Statistical model: couples $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ independently arise from the parametrized pdf

$$\underbrace{f(\boldsymbol{x}_i, \boldsymbol{z}_i)}_{\in \mathcal{F}} = \prod_{k=1}^{K} [\pi_k f_k(\boldsymbol{x}_i)]^{z_{ik}}$$

- Estimating f: implement the MLE principle through an EM-like algorithm
- Estimating K: use some information criteria as BIC, ICL,...
- Estimating $\mathbf{z}$: use the MAP principle $\hat{z}_{ik} = 1$ iif $k = \arg\max_\ell t_{i\ell}(\hat{f})$ where

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

[1] See for instance [McLachlan & Peel 2004], [Biernacki 2017]

Model-based clustering: flexibility of $\mathcal{F}$ for complex $\mathcal{X}$

- Continuous data ($\mathcal{X} = \mathbb{R}^{dx}$): multivariate Gaussian/t distrib. [McNicholas 2016]
- Categorical data: product of multinomial distributions [Goodman 1974]
- Mixing cont./cat.: product Gaussian/multinomial [Moustaki & Papageorgiou 2005]
- Functional data: the discriminative functional mixture [Bouveyron *et al.* 2015]
- Network data: the Erdös Rényi mixture [Zanghi *et al.* 2008]
- Other kinds of data, missing data, high dimension,...

Model-based clustering: poor user-friendly understanding

- n or K large: poor overview of partition $\hat{\mathbf{z}}$
- d_X large: too many parameters to embrace as a whole in $\hat{f}_k$
- Complex $\mathcal{X}$: specific and non trivial parameters involved in $\hat{f}_k$

Visualization procedures

Aim at proposing user-friendly understanding of the mathematical clustering results

Overview of clustering visualization: mapping *vs.* drawing

Visualization is the achievement of two different successive steps:

- The mapping step:
  - Performs a transformation, typically space dimension reduction of a data set or of a pdf
  - It produces no graphical output at all (deliver just a mathematical object)
- The drawing step:
  - Provides the final graphical display from the output of the previous mapping step
  - Usually involves classical graphical toolboxes and tunes any graphical parameters

Mathematician is first concerned by the more challenging mapping step

Overview of clustering visualization: individual mapping

- Aims at visualizing simultaneously the data set $\mathbf{x}$ and its estimated partition $\hat{\mathbf{z}}$
- Transforms $\mathbf{x}$, defined on $\mathcal{X}$, into $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, defined on a new space $\mathcal{Y}$

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

- Many methods, depending on $\mathcal{X}$ definition: PCA, MCA, MFA, FPCA, MDS...
- Some of them use $\hat{\mathbf{z}}$ in M^{ind}: LDA, mixture entropy preservation [Scrucca 2010]
- Nearly always, $\mathcal{Y} = \mathbb{R}^2$

Model $\hat{f}$ is is not taken into account through this approach which is focused on $\mathbf{x}$

Overview of clustering visualization: pdf mapping

- Aims at displaying information relative to the mapping of the f distribution
- Transforms $f = \sum_k \pi_k f_k \in \mathcal{F}$, into a new mixture $g = \sum_k \pi_k g_k \in \mathcal{G}$

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

- $\mathcal{G}$ is a pdf family defined on the space $\mathcal{Y}$
- M^{pdf} is often obtained as a by product of M^{ind} (tedious outside linear mappings)
- For large n, M^{ind} finally displays M^{pdf}
- Often, both $\mathbf{y}$ and g are overlaid

Summary of traditional visualization strategies[2]

Controlling the mapping family $\mathcal{M}^{\text{pdf}}$

$$\boxed{\text{Strategy}_{\mathcal{M}}}: \quad \underbrace{\mathcal{G}(\mathcal{M}^{\text{pdf}})}_{\text{uncontrolled}} = \left\{ g : g = M^{\text{pdf}}(f), f \in \mathcal{F}, M^{\text{pdf}} \in \underbrace{\mathcal{M}^{\text{pdf}}}_{\text{controlled}} \right\}$$

- Nature of $\mathcal{G}$ can dramatically depend on the choice of $\mathcal{M}^{\text{pdf}}$
- It can potentially lead to very different cluster shapes!
- Arguments for traditional $\mathcal{M}^{\text{pdf}}$: user-friendly, easy-to-compute
- Examples: linear mappings in all PCA-like methods

[2] Similar thinking with $\mathcal{M}^{\text{ind}}$

New visualization strategy

Controlling the pdf family $\mathcal{G}$

$$\boxed{\text{Strategy}_{\mathcal{G}}}: \quad \underbrace{\mathcal{M}^{\text{pdf}}(\mathcal{G})}_{\text{uncontrolled}} = \left\{ M^{\text{pdf}} : g = M^{\text{pdf}}(f), f \in \mathcal{F}, g \in \underbrace{\mathcal{G}}_{\text{controlled}} \right\}$$

- It is the reversed situation where $\mathcal{G}$ is defined instead of $\mathcal{M}^{\text{pdf}}$
- Offer opportunity to impose directly $\mathcal{G}$ to be a user-friendly mixture family
- $\text{Strategy}_{\mathcal{M}}$ and $\text{Strategy}_{\mathcal{G}}$ are both valid but $\text{Strategy}_{\mathcal{G}}$ is rarely explored!

This work: explore $\text{Strategy}_{\mathcal{G}}$

Outline

Spherical Gaussians as candidates

- Users are usually familiar with multivariate spherical Gaussians on $\mathcal{Y} = \mathbb{R}^{d_Y}$
- Thus a simple and "user-friendly" candidate g is a mixture of spherical Gaussians

$$g(\boldsymbol{y}; \boldsymbol{\mu}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{from } f} \phi_{d_Y}(\boldsymbol{y}; \underbrace{\boldsymbol{\mu}_k}_{?}, \boldsymbol{I})$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\phi_{d_Y}(.; \boldsymbol{\mu}_k, \boldsymbol{I})$ the pdf of the Gaussian distribution

- with mean $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kd_Y}) \in \mathbb{R}^{d_Y}$
- with covariance matrix equal to identity $\boldsymbol{I}$

$g(\cdot; \boldsymbol{\mu})$ should be then linked with f in order to define a sensible $\mathcal{G}$

$$\mathcal{G} = \{g : g(\cdot; \boldsymbol{\mu}), \boldsymbol{\mu} \in \arg\min \delta(f, g(\cdot; \boldsymbol{\mu})), f \in \mathcal{F}\}$$

g as the "clustering twin" of f

Question: how to choose δ since generally $\mathcal{X} \neq \mathcal{Y}$?
Answer: in our clustering context, δ should measure the clustering ability difference

Kullback-Leibler divergence of clustering ability between both f and $g(\cdot;\boldsymbol{\mu})$[3]

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

where

- p_f: pdf of proba. of classification $\mathbf{t}(f) = (\boldsymbol{t}_i(f))_{i=1}^n$, with $\boldsymbol{t}_i(f) = (t_{ik}(f))_{k=1}^{K-1}$
- $\mathsf{p}_g(\cdot;\boldsymbol{\mu})$: pdf of proba. of classif. $\mathbf{t}(g) = (\boldsymbol{t}_i(g))_{i=1}^n$, with $\boldsymbol{t}_i(g) = (t_{ik}(g))_{k=1}^{K-1}$
- $\mathcal{T} = \{\boldsymbol{t} : \boldsymbol{t} = (t_1, \ldots, t_{K-1}), t_k > 0, \sum_k t_k < 1\}$

[3] p_f is the reference measure

$\mathcal{G}$ reduced to a unique distribution

- A natural requirement: $\mathrm{p}_g(\cdot;\boldsymbol{\mu})$ and g should be linked by a one-to-one mapping
- Currently not true since rotations and/or translations are possible
- It means: for one distribution f, there is a unique optimal distribution $g(\cdot;\boldsymbol{\mu})$
- Additional constraints on $g(\cdot;\boldsymbol{\mu})$: $d_Y = K-1$, $\boldsymbol{\mu}_K = \mathbf{0}$, $\mu_{kh} = 0$ $(h > k)$, $\mu_{kk} \geq 0$

Estimating the Gaussian centers (pitch)

- The Kullback-Leibler divergence δ_{KL} has generally no closed-form
- Estimate it by the following consistent (in S) Monte-Carlo expression

$$\hat{\delta}_{\mathrm{KL}}(f, g(\cdot;\boldsymbol{\mu})) = \underbrace{\frac{1}{S}\sum_{s=1}^{S} \ln \mathrm{p}_g(\boldsymbol{t}^{(s)};\boldsymbol{\mu})}_{L(\boldsymbol{\mu};\mathbf{t})} + \mathrm{cst}$$

  with S independent draws of conditional proba. $\mathbf{t} = (\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(S)})$ from p_f
- It is the normalized (observed-data) log-likelihood function of a mixture model
- But, by construction, all the conditional probabilities are fixed in this mixture
- Thus, just maximize the normalized complete-data log-likelihood $L_{\mathrm{comp}}(\boldsymbol{\mu};\mathbf{t})$:
  - $K = 2$: this maximization is straightforward
  - $K > 2$: use a standard Quasi-Newton algorithm with different random initializations, for avoiding possible local optima

From a multivariate to a bivariate Gaussian mixture

- g is defined on $\mathbb{R}^{K-1}$ but it is more convenient to be on $\mathbb{R}^2$
- Just apply LDA on g to display this distribution on its most discriminative map
- It leads to the bivariate spherical Gaussian mixture $\tilde{g}$

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

  where $\tilde{\boldsymbol{y}} \in \mathbb{R}^2$, $\tilde{\boldsymbol{\mu}} = (\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_K)$ and $\tilde{\boldsymbol{\mu}}_k \in \mathbb{R}^2$

- Use the % of inertia of LDA to measure the quality of the mapping from g to $\tilde{g}$

Remark

If $\mathcal{X} = \mathbb{R}^d$ and f is a Gaussian mixture with isotropic covariance matrices, then the proposed mapping is equivalent to applying a LDA to the centers of f

Overall accuracy of the mapping between f and $\tilde{g}$

Use the following difference between the normalized entropies of f and $\tilde{g}$

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

- Such a quantity can be easily estimated by empirical values
- Its meaning is particularly relevant:
  - $\delta_{\mathrm{E}}(f, \tilde{g}) \approx 0$: the component overlap conveyed by $\tilde{g}$ (over f) is accurate
  - $\delta_{\mathrm{E}}(f, \tilde{g}) \approx 1$: $\tilde{g}$ strongly underestimates the component overlap of f
  - $\delta_{\mathrm{E}}(f, \tilde{g}) \approx -1$: $\tilde{g}$ strongly overestimates the component overlap of f

$\delta_{\mathrm{E}}(f, \tilde{g})$ permits to evaluate the bias of the visualization

Drawing $\tilde{g}$

- Cluster centers: the locations of $\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_K$ are materialized by vectors
- Cluster spread: the 95% confidence level displayed by a black border
- Cluster overlap: iso-probability curves of the MAP classification for different levels
- Mapping accuracy: $\delta_{\mathsf{E}}(f, \tilde{g})$ and also % of inertia by axis

Outline

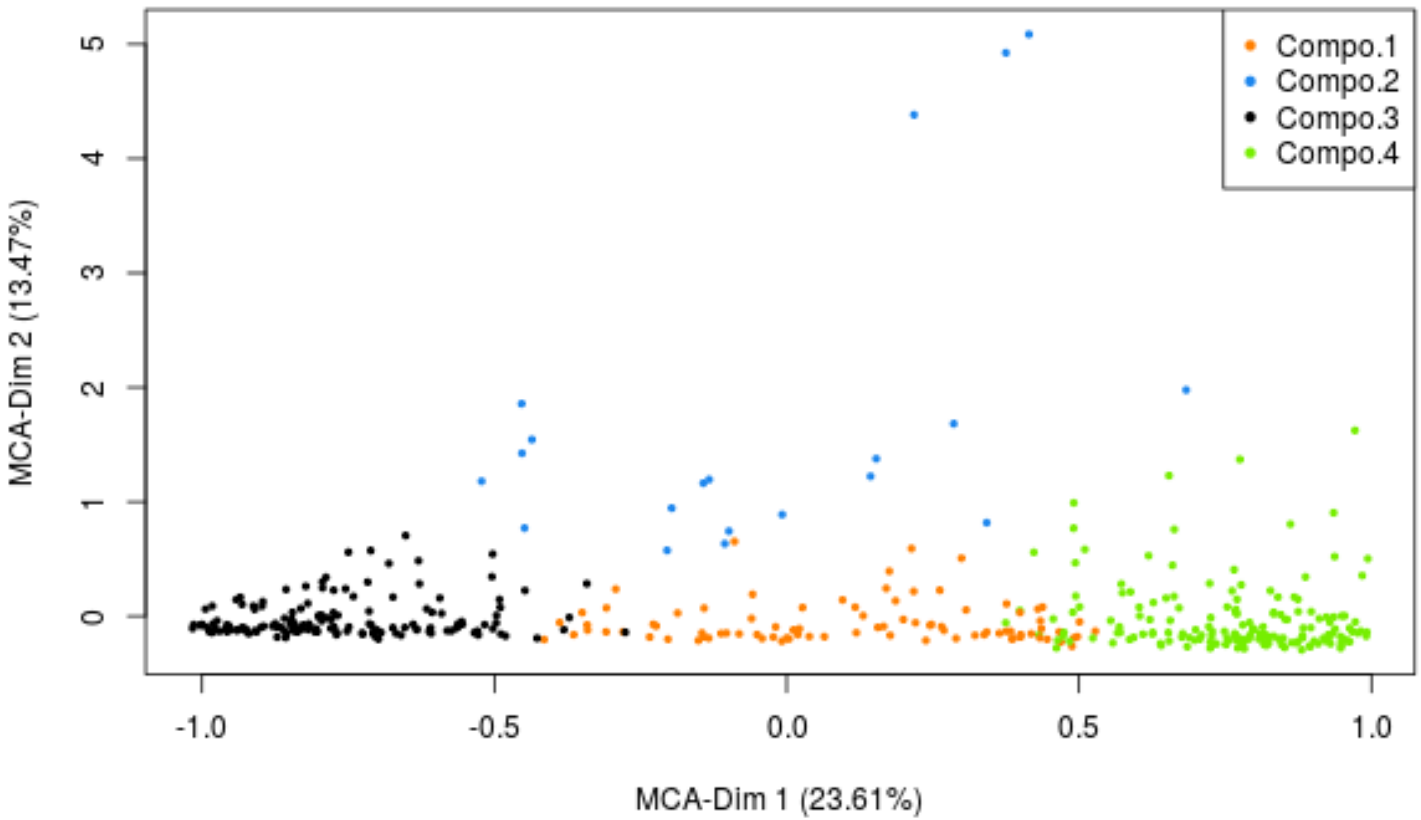
House of Representatives Congressmen: data[4] and model

- Votes of the $n = 435$ U.S. Congressmen on the $d_X = 16$ key votes
- Categorical data: for each vote, three levels are considered (yea, nay, ?)
- Data clustered by a mixture of product of multinomial distributions [Goodman 1974]
- $K = 4$ selected by BIC [Schwarz 1974]
- Use the R package `Rmixmod` [Lebret *et al.* 2015]
- Complex output: 435 individual memberships, $192 = 16 \times 3 \times 4$ parameters

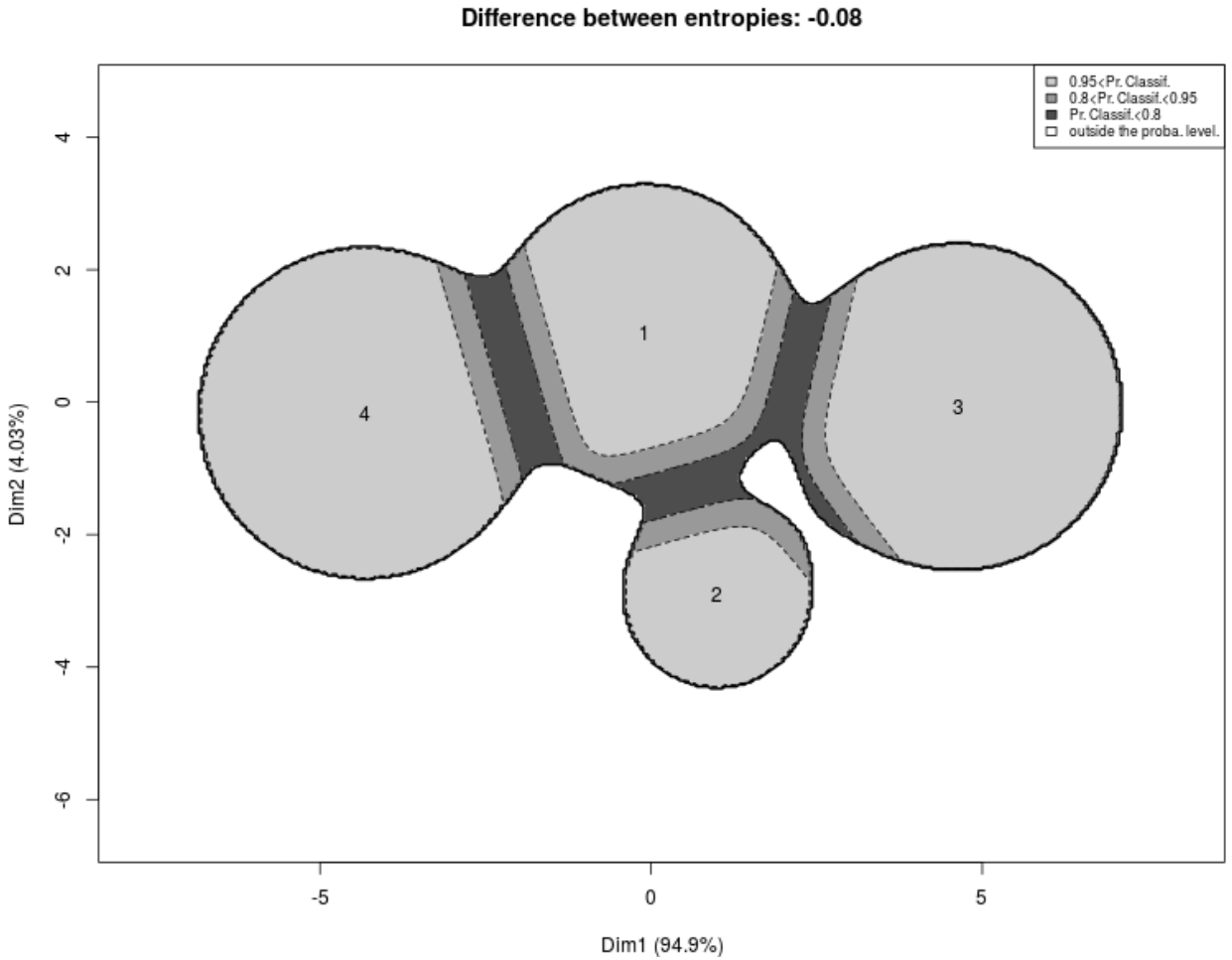
[4] [Schlimmer (1987)]

House of Representatives Congressmen: standard MCA visualization

First map of the MCA (R package `FactoMineR` [Lê *et al.* 2008]): difficult to interpret



House of Representatives Congressmen: Gaussian visualization



Mapping of f on this graph is accurate because $\delta_{\mathrm{E}}(f, \tilde{g}) = 0.01$

Contraceptive method choice: data[5] and model

- Subset of the 1987 National Indonesia Contraceptive Prevalence Survey
- Mixed data: 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)
- Clustered by a mixture f assuming that variables are independent within components
- Model selection is done by the BIC criterion which detects six components
- Use the R package `Rmixmod` [Lebret *et al.* 2015]

[5] [Lim *et al.* 2000]

Contraceptive method choice: estimated parameters

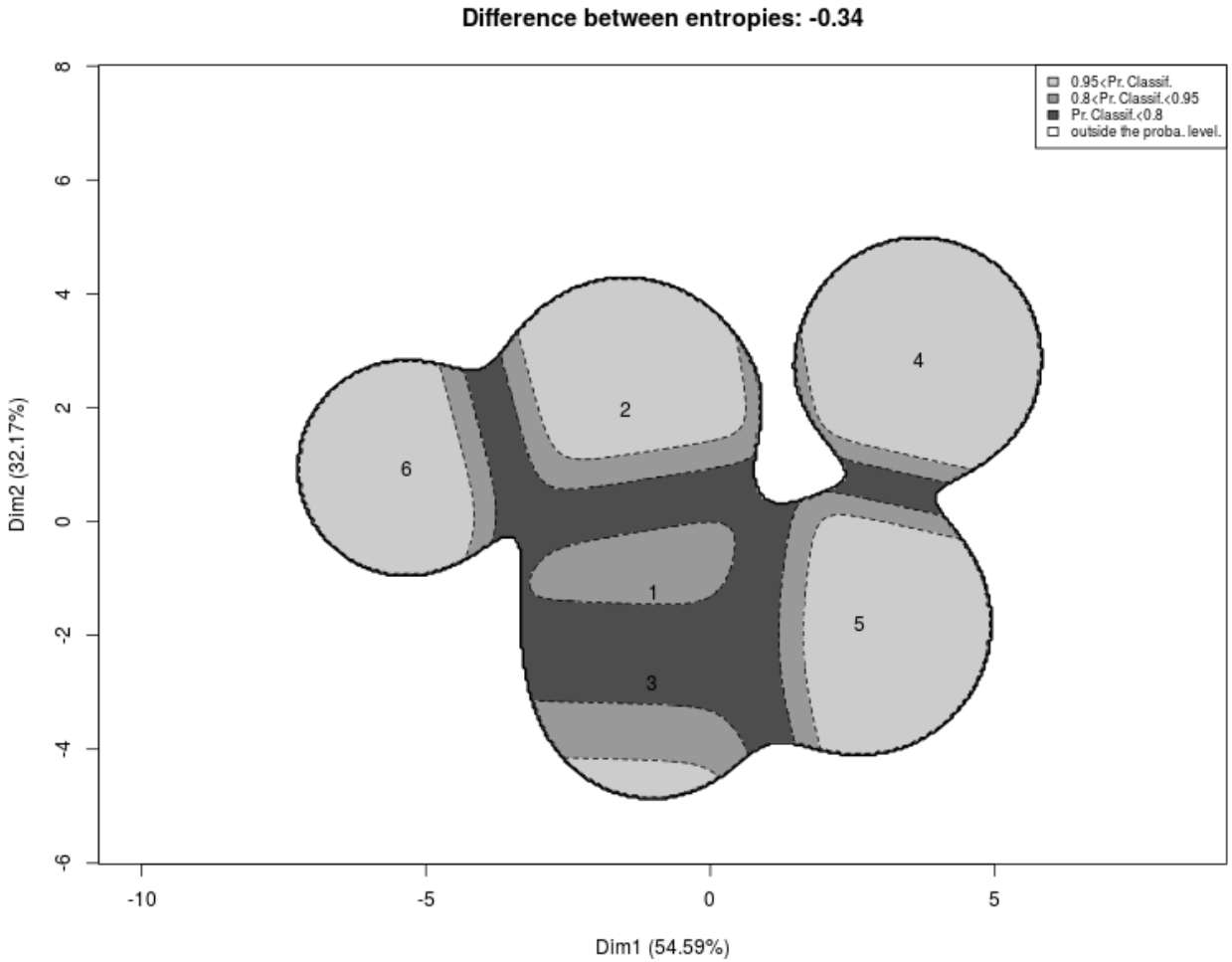
	Age		Number of children	
	Mean	Variance	Mean	Variance
Component 1	35	30	4	4
Component 2	35	22	3	2
Component 3	40	42	5	9
Component 4	25	10	1	1
Component 5	24	13	2	1
Component 6	45	7	5	8

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

Contraceptive method choice: Gaussian visualization



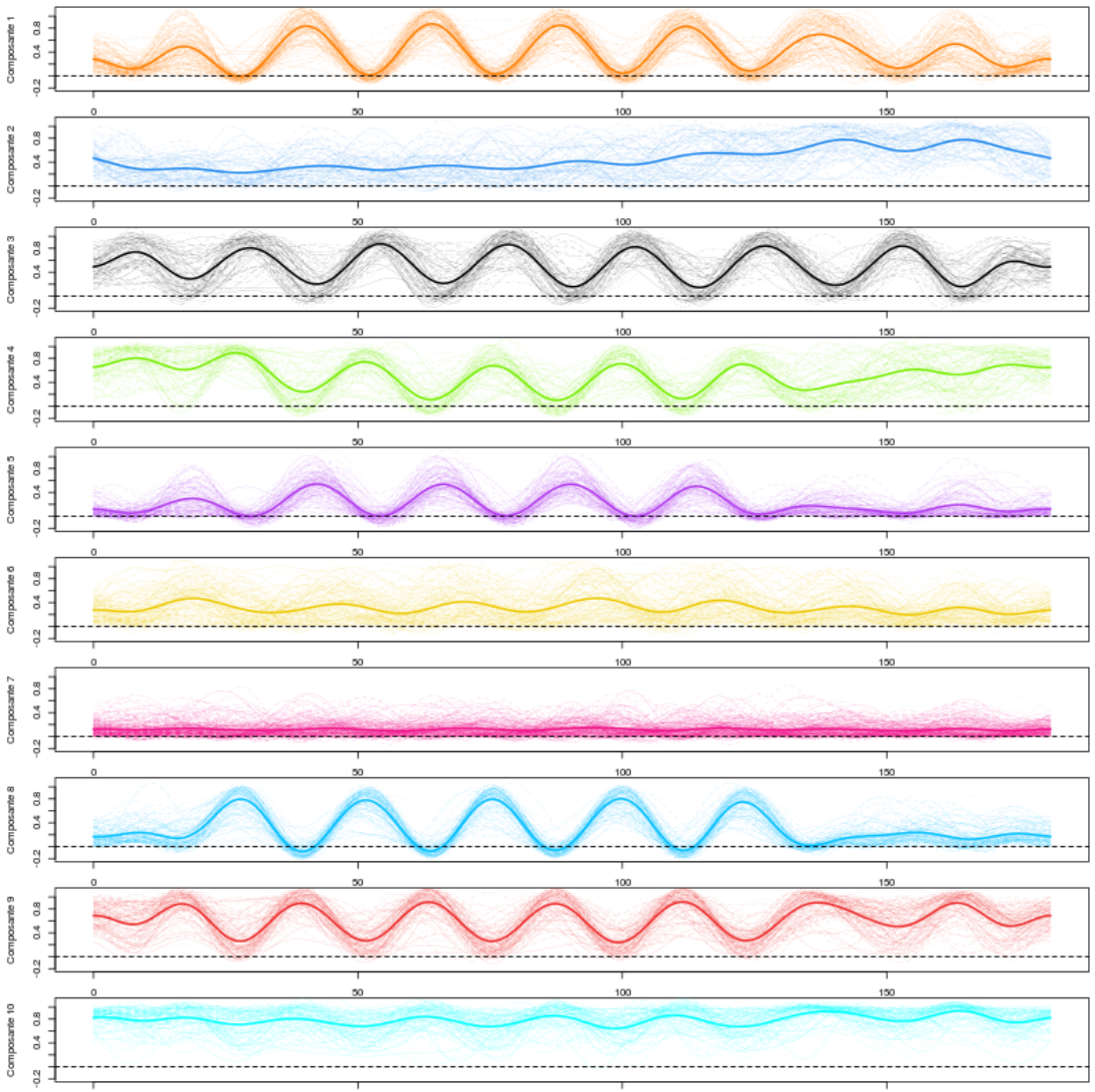
Mapping of f on this graph is accurate because $\delta_{\mathrm{E}}(f, \tilde{g}) = 0.04$

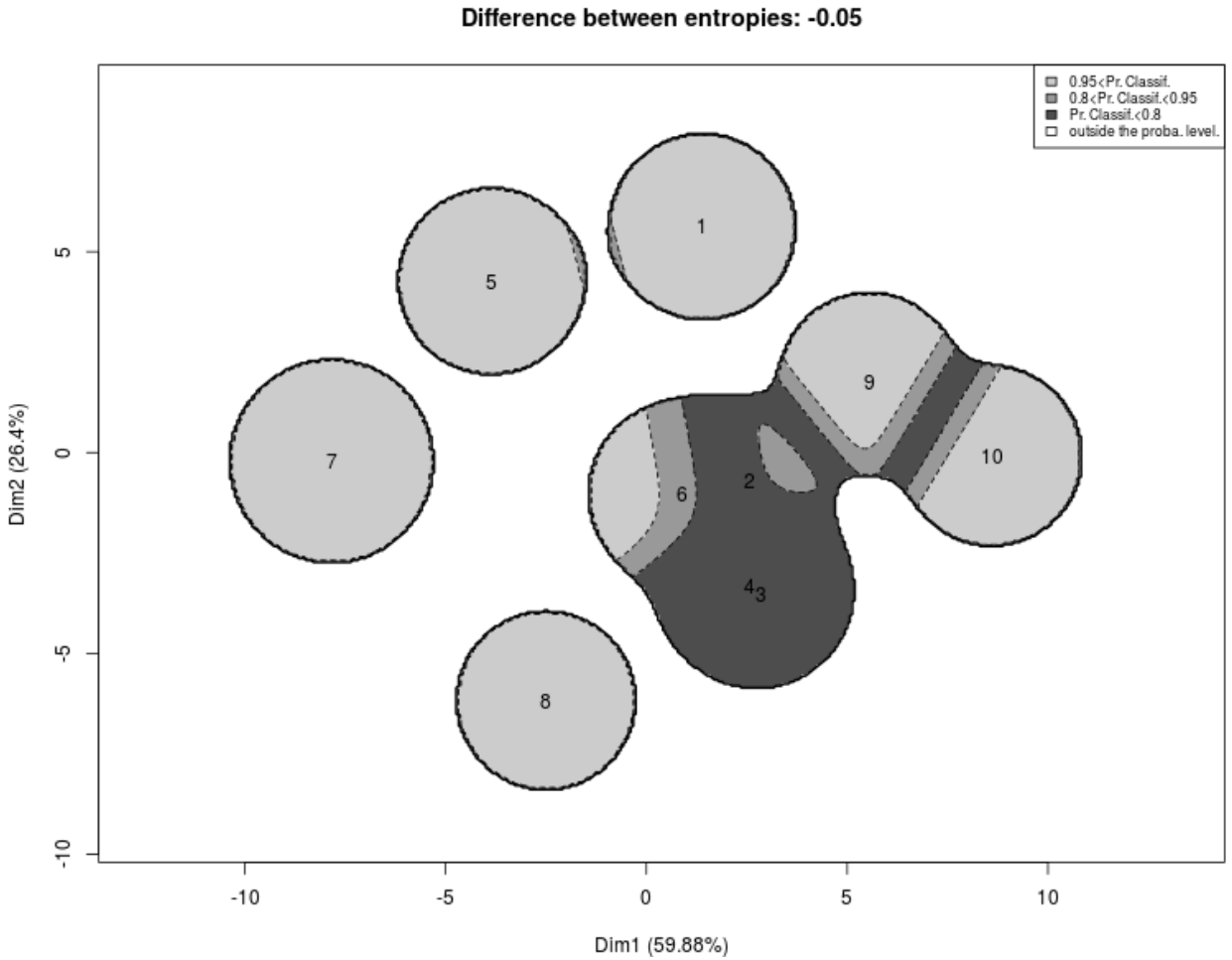
Bike sharing system: data[6] and model

- Station occupancy data collected over the course of one month on the bike sharing system in Paris
- Data collected over 5 weeks, between February, 24 and March, 30, 2014, on 1 189 bike stations
- Functional data: station status information (available bikes/docks) downloaded every hour from the open-data APIs of JCDecaux company
- The final data set contains 1 189 loading profiles, one per station, sampled at 1 448 time points
- Model: profiles of the stations were projected on a basis of 25 Fourier functions
- Model-based clustering of these functional data [Bouveyron *et al.* 2015] with the R package `FunFEM` [Bouveyron 2015]
- Retain 10 clusters

[6] [Bouveyron *et al.* (2015)]

Bike sharing system: cluster of curves visualization





Mapping of f on this graph is accurate because $\delta_{\mathrm{E}}(f, \tilde{g}) = -0.03$

French political blogosphere: data[7] and model

- Not oriented network data: a single day snapshot of over 1 100 political blogs automatically extracted the October, 14th, 2006 and manually classified by the "Observatoire Présidentielle" project.
- Nodes represent hostnames (= a set of pages) and edges represent hyperlinks between different hostnames
- Gather different communities organization due to the existence of several political parties and commentators
- Assumption: authors of these blogs tend to link, by political affinities, blogs with similar political positions
- Use the graph clustering via Erdös–Rényi mixture proposed by [Zanghi et al. 2008]
- Use the R package `mixer`
- As proposed by these authors, we consider $K = 6$ components

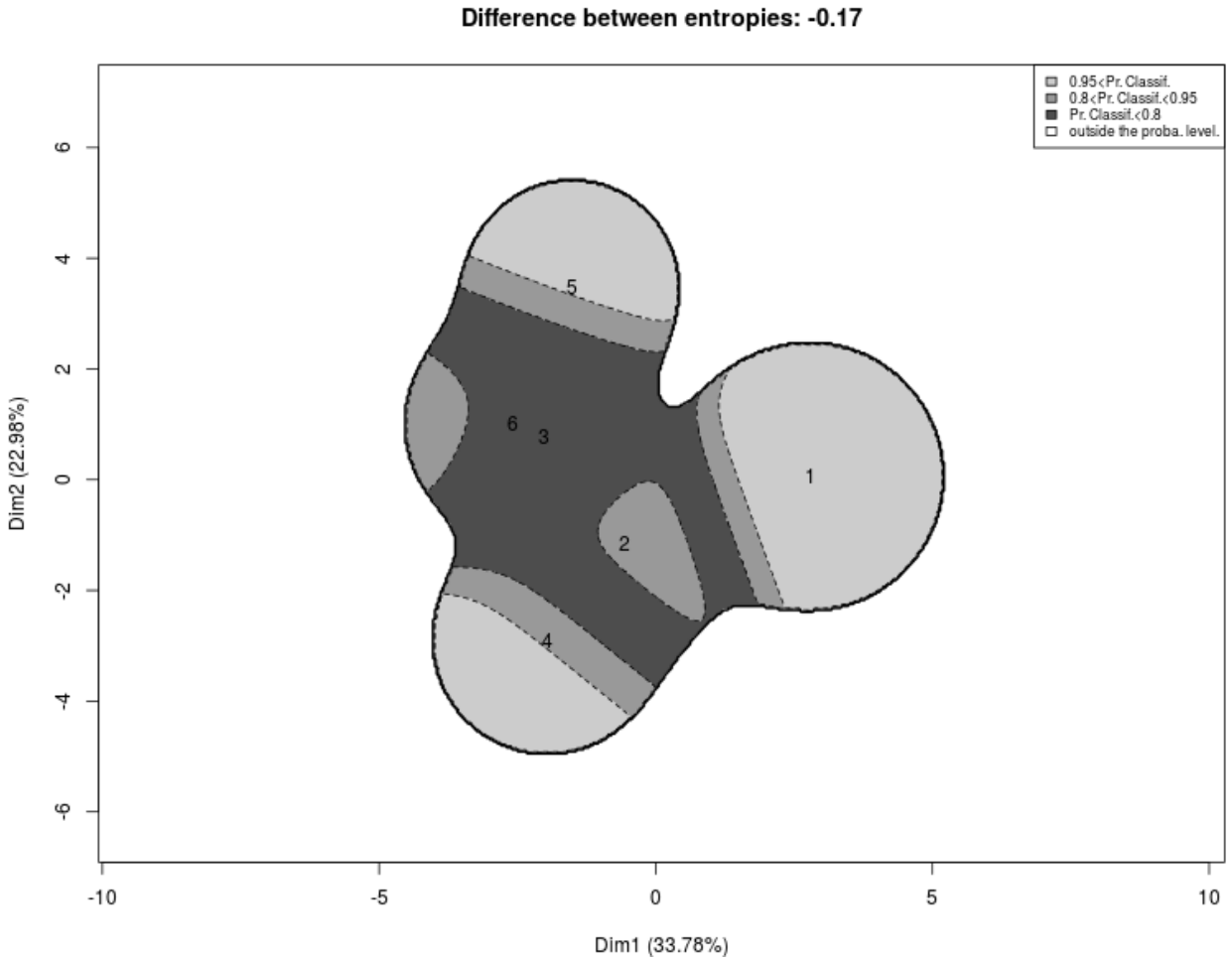
[7] [Zanghi et al. 2008]

French political blogosphere: confusion matrix

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

French political blogosphere: Gaussian visualization



The graph slightly over-represents the component overlaps: $\delta_{\mathrm{E}}(f, \tilde{g}) = -0.216$

Outline

Correspondence between new axis 1[8] and initial features

We measure the overlap strength of each feature j by the following entropy ratio:

$$\bar{E}_j = \frac{E_j}{\sum_{j=1}^{d} E_j} \qquad \in [0,1]$$

where E_j corresponds to the marginal entropy of axis 1 re-estimated without feature j

- The larger is $\bar{E}_j$, the more discriminant is feature j on axis 1
- Since axis 1 is the most discriminant axis, $\bar{E}_j$ measures the contribution of feature j to the mixture overlap on axis 1

[8] The same principle could be applied to all other new axes.

Prostate cancer data[10]: definition

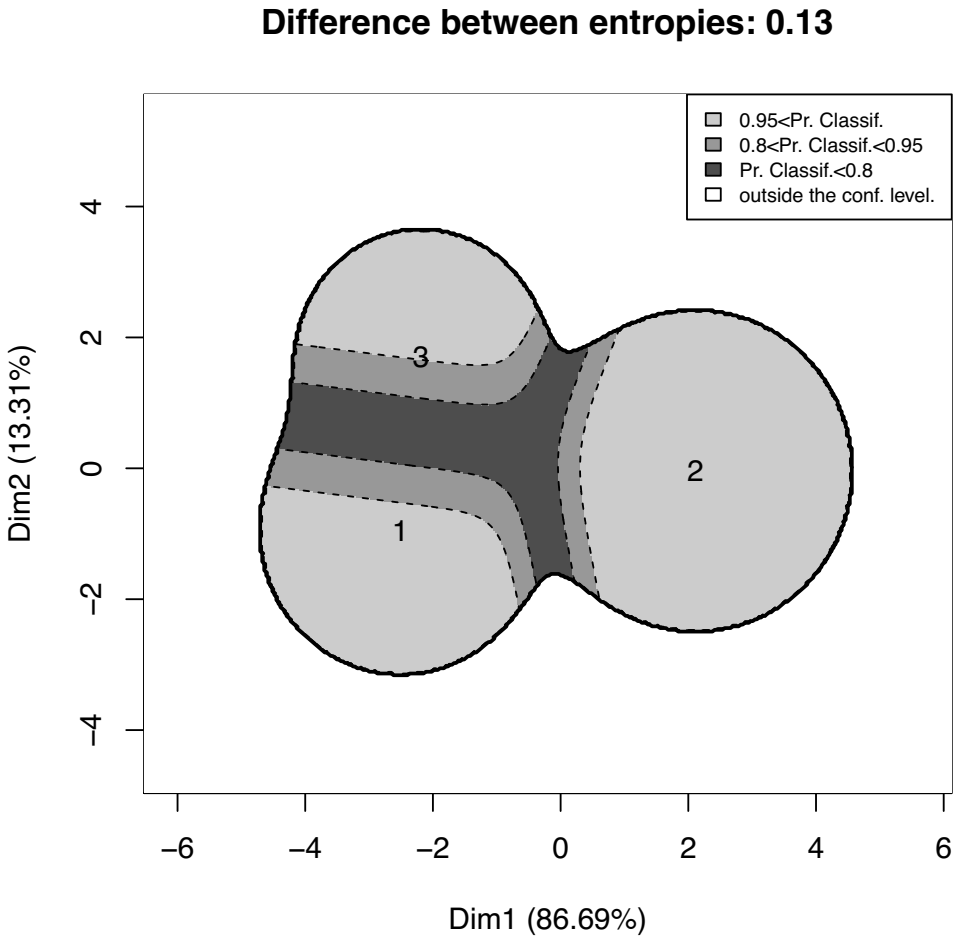
- Individuals: 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Some missing data: 62 missing values ($\approx 1\%$)

Use the Rmixtcomp package[9] to estimate the mixture...

[9] https://cran.r-project.org/web/packages/RMixtComp/index.html

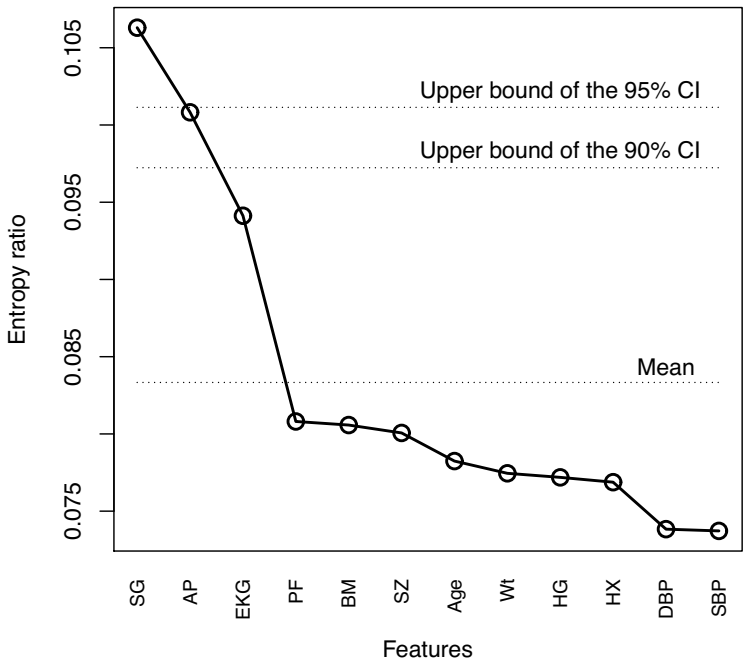
[10] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Prostate cancer data: visualization[11]



[11] Biernacki, C., Marbac, M., Vandewalle, V. Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering. J Classif (2020).

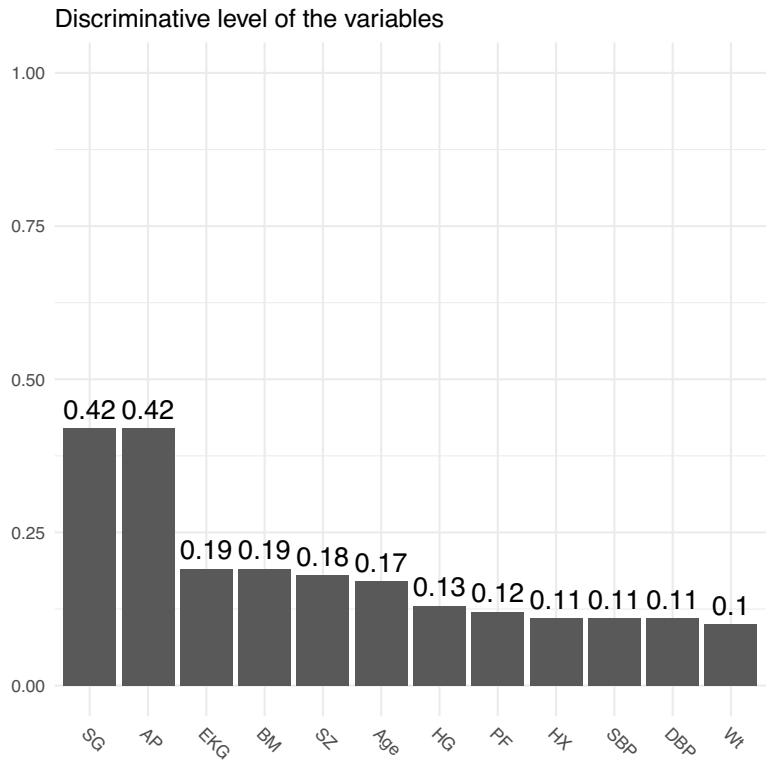
Prostate cancer data: axis 1 interpretation



- Features SG (and AP) are the most discriminant features
- Thus axis 1 can be interpreted as a "SG/AP axis"

Prostate cancer data: validation of axis 1 interpretation

In the Rmixcomp package, there exists an estimation of the feature importance (which is independent of the Gaussian-based visualization)



We remark that Rmixcomp feature importance and visualization interpretation are in accordance. . .

Outline

Conclusion

- Generic method for visualizing the results of any model-based clustering result
- Very easy to understand output since "Gaussian-like"
- Permit visualization for any type of data, because only based on proba. of classif.
- Can be used after any existing package of model-based clustering
- The overall accuracy of the visualization is also provided

R package on the CRAN: `ClusVis`

Extensions

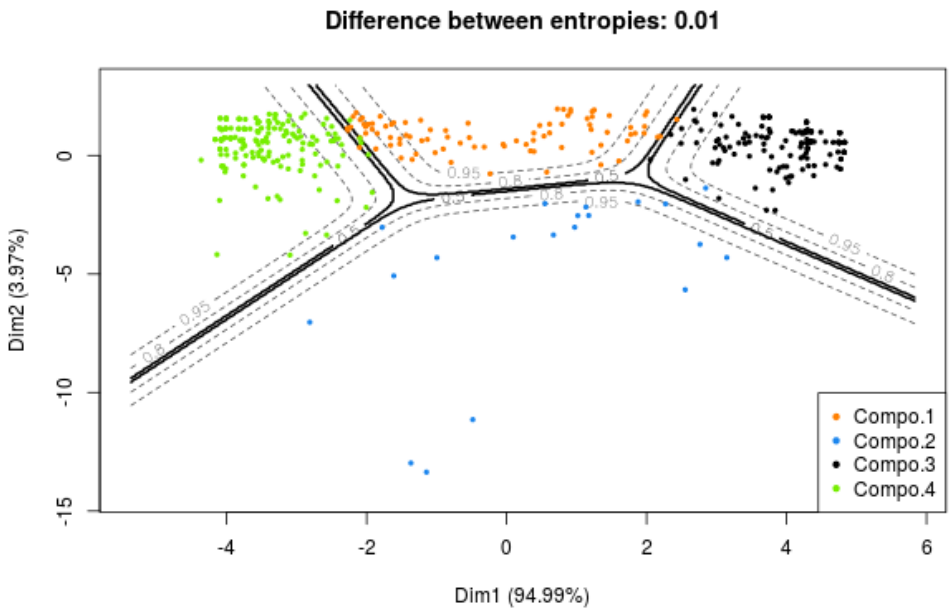
- Possibility to explore other pdf visualizations than Gaussians
- However, should keep in mind simple visualizations are targeted
- Possibility to compare/select pdf candidates through δ_{KL} or δ_{E}

About individual visualization

- Theoretically, impossible to obtain individual visualization from pdf visualization
- However, we can propose a pseudo scatter plot of $\mathbf{x}$ as follows

$$\boldsymbol{x}_i \longmapsto \boldsymbol{t}_i(f) = \boldsymbol{t}_i(g) \overset{\text{bijection}}{\longmapsto} \boldsymbol{y}_i \in \mathbb{R}^{K-1} \overset{\text{LDA}}{\longmapsto} \tilde{\boldsymbol{y}}_i \in \mathbb{R}^2$$

- $\tilde{\mathbf{y}}$ allows only to visualize the classification position of $\mathbf{x}$
- Example for the congressmen data set



- Caution: do not overlay pdf and individual plots since $\tilde{\mathbf{y}} = (\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n)$ is not necessarily drawn from a Gaussian mixture