



**HAL**  
open science

# Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle

► **To cite this version:**

Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle. Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering. SPSR 2019, Apr 2019, Bucarest, Romania. hal-02400486

**HAL Id: hal-02400486**

**<https://hal.science/hal-02400486>**

Submitted on 9 Dec 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

M. Marbac-Lourdelle, C. Biernacki, V. Vandewalle

SPSR 2019

10-11 May 2019, Bucharest, Romania

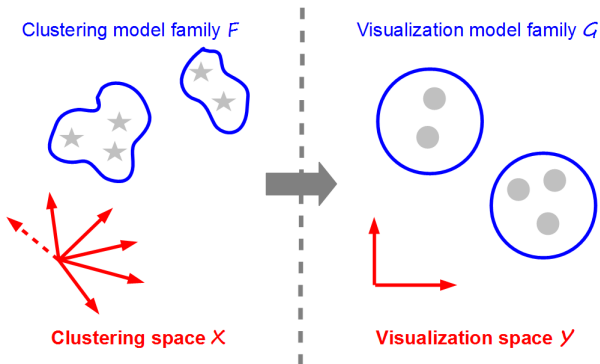


## Take home message

### Data/density visualization

**Traditionally:** chose the form of the mapping from  $\mathcal{X}$  to  $\mathcal{Y}$  for *user convenience*

**Proposal:** chose the form of the density of the data on  $\mathcal{Y}$  for *cluster interpretation convenience*



# Outline

- 1 Clustering: from modeling to visualizing
- 2 Mapping clusters as spherical Gaussians
- 3 Numerical illustrations for functional data
- 4 Discussion

## Model-based clustering: pitch<sup>1</sup>

- **Data set:**  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ , each  $\mathbf{x}_i \in \mathcal{X}$  with  $d_X$  variables (possibly mixing continuous, categorical, functional...)
- **Unknown partition in  $K$  clusters:**  $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$  with binary notation  $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$
- **Statistical model:** couples  $(\mathbf{x}_i, \mathbf{z}_i)$  independently arise from the parametrized pdf

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

- **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  $\hat{z}_{ik} = 1$  iff  $k = \arg \max_{\ell} t_{i\ell}(\hat{f})$  where

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

<sup>1</sup>See for instance [McLachlan & Peel 2004], [Biernacki 2017]

## Model-based clustering: poor user-friendly understanding

- $n$  or  $K$  large: poor overview of partition  $\hat{z}$
- $d_X$  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: individual and pdf mappings

### Individual mapping: visualize $\mathbf{x}$ and its estimated partition $\hat{\mathbf{z}}$

- Transforms  $\mathbf{x}$ , defined on  $\mathcal{X}$ , into  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ , 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  $M^{\text{ind}}$ : LDA, mixture entropy preservation [Scrucca 2010]
- Nearly always,  $\mathcal{Y} = \mathbb{R}^2$
- Model  $\hat{f}(\mathbf{x}, \mathbf{z})$  is not taken into account, approach focused on  $\mathbf{x}$

## Overview of clustering visualization: individual and pdf mappings

### Individual mapping: visualize $\mathbf{x}$ and its estimated partition $\hat{\mathbf{z}}$

- Transforms  $\mathbf{x}$ , defined on  $\mathcal{X}$ , into  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ , 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  $M^{\text{ind}}$ : LDA, mixture entropy preservation [Scrucca 2010]
- Nearly always,  $\mathcal{Y} = \mathbb{R}^2$
- Model  $\hat{f}(\mathbf{x}, \mathbf{z})$  is not taken into account, approach focused on  $\mathbf{x}$

### Pdf mapping: display information relative to 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  $\mathcal{Y}$
- $M^{\text{pdf}}$  is often obtained as a by product of  $M^{\text{ind}}$  (variable change formula)
- $\mathcal{G}$  is not a usual mixture family (Gaussian, ...) when  $M^{\text{ind}}$  is a nonlinear mapping
- For large  $n$ ,  $M^{\text{ind}}$  finally displays  $M^{\text{pdf}}$
- Often, both  $\mathbf{y}$  and  $g$  are overlaid



## Traditional visualization strategies and proposal

Traditional strategies: Controlling the mapping family  $\mathcal{M}^{\text{pdf}}$ <sup>2</sup>

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

- 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

---

<sup>2</sup>Similar thinking with  $\mathcal{M}^{\text{ind}}$

## Traditional visualization strategies and proposal

Traditional strategies: Controlling the mapping family  $\mathcal{M}^{\text{pdf}}$ <sup>2</sup>

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

- 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

Proposed strategy: Controlling the pdf family  $\mathcal{G}$

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

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

<sup>2</sup>Similar thinking with  $\mathcal{M}^{\text{ind}}$

# Outline

- 1 Clustering: from modeling to visualizing
- 2 Mapping clusters as spherical Gaussians**
- 3 Numerical illustrations for functional data
- 4 Discussion

## 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

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

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

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

$g(\cdot; \boldsymbol{\mu})$  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  $\delta$  since generally  $\mathcal{X} \neq \mathcal{Y}$ ?

**Answer:** in our clustering context,  $\delta$  should measure the **clustering ability difference**

Kullback-Leibler divergence of clustering ability between both  $f$  and  $g(\cdot; \boldsymbol{\mu})^2$

$$\delta_{\text{KL}}(f, g(\cdot; \boldsymbol{\mu})) = \int_{\mathcal{T}} p_f(\mathbf{t}) \ln \frac{p_f(\mathbf{t})}{p_g(\mathbf{t}; \boldsymbol{\mu})} d\mathbf{t}$$

where

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

**Thus**,  $g$  should produce a distribution of the class membership posterior probabilities similar the one resulting of  $f$ .

---

<sup>2</sup> $p_f$  is the reference measure

## $g$ as the “clustering twin” of $f$

**Question:** how to choose  $\delta$  since generally  $\mathcal{X} \neq \mathcal{Y}$ ?

**Answer:** in our clustering context,  $\delta$  should measure the **clustering ability difference**

Kullback-Leibler divergence of clustering ability between both  $f$  and  $g(\cdot; \boldsymbol{\mu})^2$

$$\delta_{\text{KL}}(f, g(\cdot; \boldsymbol{\mu})) = \int_{\mathcal{T}} p_f(\mathbf{t}) \ln \frac{p_f(\mathbf{t})}{p_g(\mathbf{t}; \boldsymbol{\mu})} d\mathbf{t}$$

where

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

**Thus**,  $g$  should produce a distribution of the class membership posterior probabilities similar the one resulting of  $f$ .

- A natural requirement:  $p_g(\cdot; \boldsymbol{\mu})$  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  $g(\cdot; \boldsymbol{\mu})$ :  $d\gamma = K - 1$ ,  $\boldsymbol{\mu}_K = \mathbf{0}$ ,  $\mu_{kh} = 0$  ( $h > k$ ),  $\mu_{kk} \geq 0$

<sup>2</sup> $p_f$  is the reference measure

## Estimating the Gaussian centers

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

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

with  $S$  independent draws of conditional proba.  $\mathbf{t} = (\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(S)})$  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  $L_{\text{comp}}(\boldsymbol{\mu}; \mathbf{t})$ :
  - $K = 2$ : 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  $\tilde{g}$

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

where  $\tilde{\mathbf{y}} \in \mathbb{R}^2$ ,  $\tilde{\boldsymbol{\mu}} = (\tilde{\boldsymbol{\mu}}_1, \dots, \tilde{\boldsymbol{\mu}}_K)$  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  $\tilde{g}$

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

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

$\delta_E(f, \tilde{g})$  permits to **evaluate the bias of the visualization**

## Drawing $\tilde{g}$

- **Cluster centers:** the locations of  $\tilde{\mu}_1, \dots, \tilde{\mu}_K$  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:**  $\delta_E(f, \tilde{g})$  and also % of inertia by axis

# Outline

- 1 Clustering: from modeling to visualizing
- 2 Mapping clusters as spherical Gaussians
- 3 Numerical illustrations for functional data**
- 4 Discussion

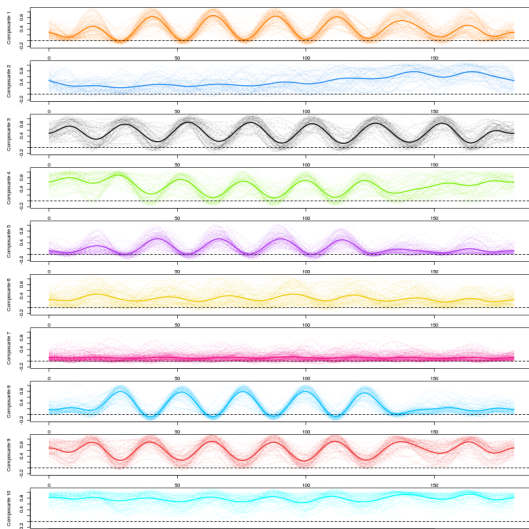
## Bike sharing system: data<sup>3</sup> and model

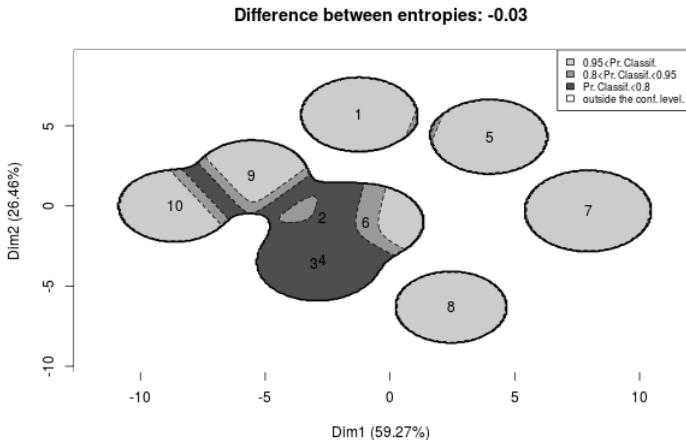
- Station occupancy data collected over the course of one month on the bike sharing system in Paris
- Data collected over 5 weeks, between February, 24 and March, 30, 2014, on 1 189 bike stations
- **Functional data**: station status information (available bikes/docks) downloaded every hour from the 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
- Visualization using ClusVis R package

---

<sup>3</sup>[Bouveyron *et al.* (2015)]

## Bike sharing system: cluster of curves visualization





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

# Outline

- 1 Clustering: from modeling to visualizing
- 2 Mapping clusters as spherical Gaussians
- 3 Numerical illustrations for functional data
- 4 Discussion**

## Conclusion and extensions

### Conclusion

- Generic method for visualizing the results of a model-based clustering
- Very easy to understand output since “Gaussian-like”
- Permits 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

### Extensions

- Possibility to explore other pdf visualizations than Gaussians
- However, should keep in mind simple visualizations are targeted
- Possibility to compare pdf candidates through  $\delta_{KL}$  or  $\delta_E$

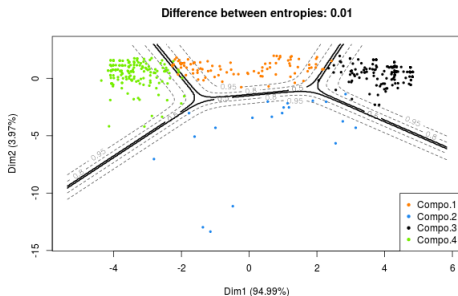


## About individual visualization

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

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

- $\tilde{\mathbf{y}}$  allows only to visualize the classification position of  $\mathbf{x}$



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

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

- **Continuous data** ( $\mathcal{X} = \mathbb{R}^{d_x}$ ): 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 Erdős Rényi mixture [Zanghi *et al.* 2008]
- Other kinds of data, missing data, high dimension, . . .