



HAL
open science

Traitement statistique des données manquantes-Part II Numerical and non-numerical data

Christophe Biernacki

► **To cite this version:**

Christophe Biernacki. Traitement statistique des données manquantes-Part II Numerical and non-numerical data. Doctorat. France. 2021. hal-03505650

HAL Id: hal-03505650

<https://hal.science/hal-03505650>

Submitted on 31 Dec 2021

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.

Traitement statistique des données manquantes

Atelier Statistique de la SFdS – 10 et 11 mars 2021

—

Part II Numerical and non-numerical data

C. Biernacki

Laboratoire P. Painlevé, UMR CNRS 8524 & Université de Lille & Inria



Merci à : F. Antonazzo, C. Boyer, G. Celeux, Q. Grimonprez, J. Jacques, J. Josse,
C. Keribin, V. Kubicki, F. Laporte, M. Marbac, A. Sportisse, J. Vandaele, V. Vandewalle

Outline

1 Mixture models as a multi-purpose tool

2 Gaussian case

3 Mixed data case

4 Ranking data case

5 RMixtComp in practice

6 Rankcluster in practice

Part I argued in favor of modeling to deal with missing data

Mixture models are a generic/flexible modeling for addressing many classical statistical purposes

Parametric mixture model

- Unknown true distribution¹:

$$\mathbf{x} = (\mathbf{x}^o, \mathbf{x}^m) = (\mathbf{x}_1, \dots, \mathbf{x}_n) \stackrel{iid}{\sim} p(\cdot)$$

- Parametric mixture assumption:

$$p(\mathbf{x}_1) = p(\mathbf{x}_1; \boldsymbol{\theta}) = \sum_{k=1}^K \underbrace{\pi_k}_{\text{proportion}} \underbrace{p(\mathbf{x}_1; \boldsymbol{\alpha}_k)}_{\text{component}}$$

- Mixture parameter:

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}) \text{ with } \boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \text{ and } \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K)$$

- Model: it includes both the family $p(\cdot; \boldsymbol{\alpha}_k)$ and the number of components K

$$\mathbf{m} = \{p(\mathbf{x}_1; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

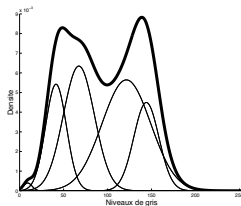
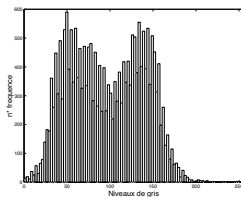
The number of free *continuous* parameters is given by

$$\nu = \dim(\Theta)$$

¹See Part I for for precise definition of \mathbf{x}^o and \mathbf{x}^m .

Illustration of mixture models flexibility (1/2)

- **Mixture models:** extremely flexible family of distributions



- **Mixture of mixture models:** flexibility for groups also

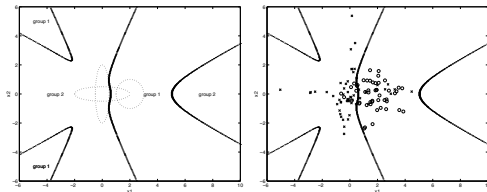
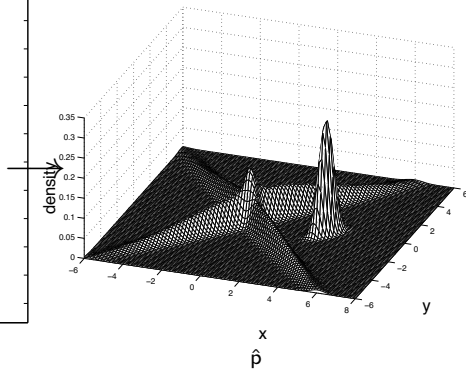
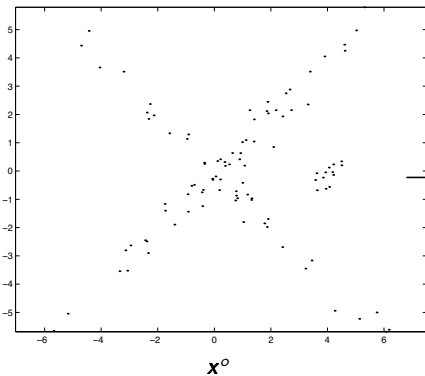


Illustration of mixture models flexibility (2/2)



In many situations, it can be theoretically proved that if K is large enough then a mixture distribution can approximate any distribution!

Sampling assumptions as a missing variable formulation

- A mixture model can be expressed through a **binary latent variable**:

$$\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$$

where $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ with $z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^K z_{ik} = 1$

- **Generative process**:

$$\begin{aligned} \mathbf{z}_1, \dots, \mathbf{z}_n &\stackrel{iid}{\sim} \text{Mult}_K(\mathbf{1}, \boldsymbol{\pi}) \\ \mathbf{x}_i |_{z_{ik}=1} &\stackrel{ind}{\sim} p(\cdot; \boldsymbol{\alpha}_k) \end{aligned}$$

- **Joint and marginal (or mixture) distributions**:

$$p(\mathbf{x}_1, \mathbf{z}_1) = \prod_{k=1}^K [\pi_k p(\mathbf{x}_1; \boldsymbol{\alpha}_k)]^{z_{1k}} \quad , \quad p(\mathbf{x}_1) = \sum_{k=1}^K \pi_k p(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

With this latent variable formulation, many applications are possible. . .

The mixture model answer for imputation

Straightforward consequence of the fact that $p(\cdot; \theta)$ is flexible enough to approximate any p

- **Single imputation:** Straightforward, for instance by the mode²

$$\hat{\mathbf{x}}^m = \arg \max_{\mathbf{x}^m} p(\mathbf{x}^m | \mathbf{x}^o; \theta)$$

- **Multiple imputation:** draw multiple values of \mathbf{x}^m from the distribution $p(\mathbf{x}^m | \mathbf{x}^o; \theta)$

²Other possibilities, depending on the data type: mean, etc.

The mixture model answer for supervised classification (1/2)

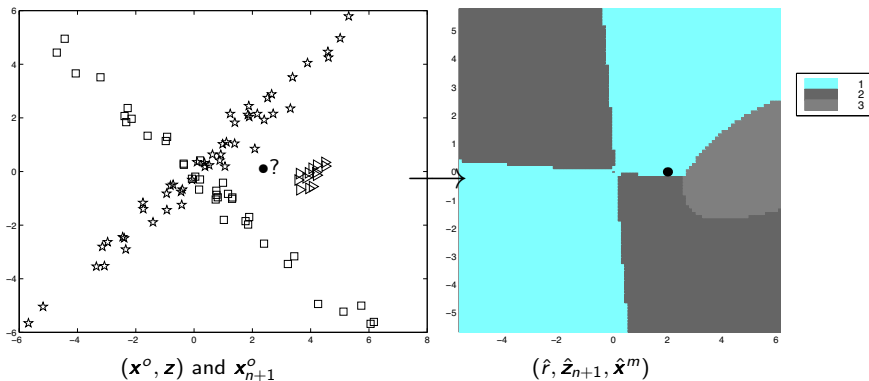
Aim: estimation of an allocation rule $r(\cdot; \theta)$ from $(\mathbf{x}^o, \mathbf{z}^o)$

$$r(\cdot; \theta) : \begin{array}{l} \mathcal{X} \longrightarrow \{1, \dots, K\} \\ \mathbf{x}_{n+1} \longmapsto r(\mathbf{x}_{n+1}; \theta) \end{array}$$

Mixed, missing, uncertain

| Individuals \mathbf{x}^o | | | | Partition \mathbf{z} | | | \Leftrightarrow | Group |
|----------------------------|-------------|--------------|---------|------------------------|---|---|-------------------|-------|
| ? | 0.5 | red | 5 | 0 | 1 | 0 | \Leftrightarrow | G_2 |
| 0.3 | 0.1 | green | 3 | 1 | 0 | 0 | \Leftrightarrow | G_1 |
| 0.3 | 0.6 | {red, green} | 3 | 1 | 0 | 0 | \Leftrightarrow | G_1 |
| 0.9 | [0.25 0.45] | red | ? | 0 | 0 | 1 | \Leftrightarrow | G_3 |
| ↓ | ↓ | ↓ | ↓ | | | | | |
| continuous | continuous | categorical | integer | | | | | |

The mixture model answer for supervised classification (2/2)



The mixture model answer for semi-supervised classification (1/2)

- **Aim:** estimation of an allocation rule $r(\cdot; \theta)$ from $(\mathbf{x}^o, \mathbf{z}^o)$

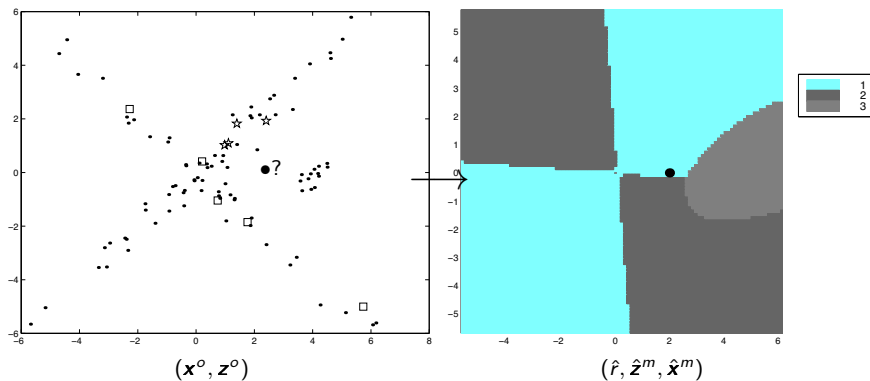
$$r(\cdot; \theta) : \begin{array}{l} \mathcal{X} \\ \mathbf{x}_{n+1} \end{array} \begin{array}{l} \longrightarrow \\ \longmapsto \end{array} \begin{array}{l} \{1, \dots, K\} \\ r(\mathbf{x}_{n+1}; \theta) \end{array}$$

- Happen when \mathbf{x} is cheaper than \mathbf{z}

Mixed, missing, uncertain

| | Individuals \mathbf{x}^o | | | | Partition \mathbf{z}^o | | | \Leftrightarrow | Group |
|------------|----------------------------|--------------|---------|---|--------------------------|---|-------------------|-------------------|-------|
| ? | 0.5 | red | 5 | 0 | ? | ? | \Leftrightarrow | G_2 or G_3 | |
| 0.3 | 0.1 | green | 3 | 1 | 0 | 0 | \Leftrightarrow | G_1 | |
| 0.3 | 0.6 | {red, green} | 3 | ? | ? | ? | \Leftrightarrow | ??? | |
| 0.9 | [0.25 0.45] | red | ? | 0 | 0 | 1 | \Leftrightarrow | G_3 | |
| ↓ | ↓ | ↓ | ↓ | | | | | | |
| continuous | continuous | categorical | integer | | | | | | |

The mixture model answer for semi-supervised classification (2/2)



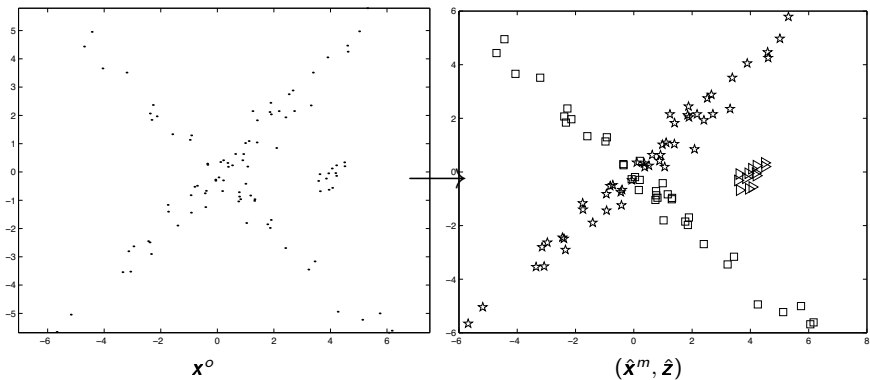
The mixture model answer for unsupervised classification (1/2)

Aim: estimation of $\mathbf{z}^m = \mathbf{z}$ from \mathbf{x}^o

Mixed, missing, uncertain

| | Individuals \mathbf{x}^o | | | | Partition \mathbf{z} | \Leftrightarrow | Clusters |
|------------|----------------------------|-------------|---------|-------|------------------------|-------------------|----------|
| ? | 0.5 | red | 5 | ? ? ? | \Leftrightarrow | ??? | |
| 0.3 | 0.1 | green | 3 | ? ? ? | \Leftrightarrow | ??? | |
| 0.3 | 0.6 | {red,green} | 3 | ? ? ? | \Leftrightarrow | ??? | |
| 0.9 | [0.25 0.45] | red | ? | ? ? ? | \Leftrightarrow | ??? | |
| ↓ | ↓ | ↓ | ↓ | | | | |
| continuous | continuous | categorical | integer | | | | |

The mixture model answer for unsupervised classification (2/2)



The mixture model answer in $\{\emptyset, \text{semi}, \text{un}\}$ classification

- Rigorous definition of a group:

$$\mathbf{x}_1 \in G_k \Leftrightarrow z_{1k} = 1$$

- Maximum *a posteriori* (MAP):

$$t_k(\mathbf{x}_1; \theta) = p(z_{1k} = 1 | \mathbf{x}_1) = \frac{\pi_k p(\mathbf{x}_1; \alpha_k)}{p(\mathbf{x}_1; \theta)}$$

$$r(\mathbf{x}_1; \theta) = \arg \max_{k=\{1, \dots, K\}} t_k(\mathbf{x}_1; \theta)$$

$$\hat{z}_{1r(\mathbf{x}_1; \theta)} = 1$$

The central question is now to estimate θ . We use the maximum observed log-likelihood principle under the MAR principle:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \ell(\theta; \mathbf{x}^\circ, \mathbf{z}^\circ)$$

EM and SEM algorithms

EM

- Initialisation: $\theta^{(0)}$
- Iteration (q):
 - **E-step**: compute $Q(\theta, \theta^{(q)}) = E[\ell_c(\theta; \mathbf{x}^o, \mathbf{x}^m, \mathbf{z}^o, \mathbf{z}^m) | \mathbf{x}^o, \mathbf{z}^o; \theta^{(q)}]$
 - **M-step**: maximize $\theta^{(q+1)} = \arg \max_{\theta} Q(\theta, \theta^{(q)})$
- Stopping rule: iteration number Q or criterion stability

SEM

- Initialisation: $\theta^{(0)}$
- Iteration (q):
 - **SE-step**: draw $(\mathbf{x}^{m(q)}, \mathbf{z}^{m(q)})$ from $p(\mathbf{x}^m, \mathbf{z}^o | \mathbf{x}^o, \mathbf{z}^o; \theta^{(q)})$
 - **M-step**: maximize $\theta^{(q+1)} = \arg \max_{\theta} \ell_c(\theta; \mathbf{x}^o, \mathbf{x}^{m(q)}, \mathbf{z}^o, \mathbf{z}^{m(q)})$
- Stopping rule: iteration number Q

See Part I for more information...

Theoretical model selection criteria

The most widespread principle

$$\underbrace{\text{Criterion}}_{\text{to be maximized}} = \underbrace{\text{maximum observed log-likelihood}}_{\text{model-data adequacy}} - \underbrace{\text{penalty}}_{\text{"cost" of the model}}$$

| crit erion | pen alty | inter pretation | user pur pose |
|-------------------------|-----------------------|------------------------------|----------------------------|
|-------------------------|-----------------------|------------------------------|----------------------------|

general criteria in statistics

| | | | |
|-----|-----------------|------------------|--------------------|
| AIC | ν | model complexity | (semi-)supervised |
| BIC | $0.5\nu \ln(n)$ | model complexity | density estimation |

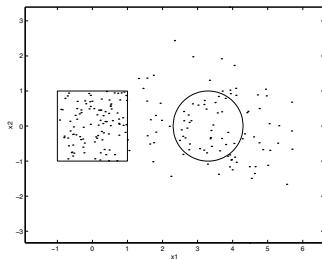
specific criterion for the clustering aim

| | | | |
|-----|--|---|------------|
| ICL | $0.5\nu \ln(n) - \sum_{i,k} \hat{z}_{ik} \ln t_{ik}(\hat{\theta})$ | model complexity + partition entropy | clustering |
|-----|--|---|------------|

N.B.: in a (semi-) supervised context, it is also possible to use the predictive error rate

The ICL criterion: robustness to model misspecification

- A bivariate mixture of a uniform and a Gaussian cluster:
 - non-Gaussian component: $\pi_1 = 0.5$, $p_1(\mathbf{x}_1) = 0.25 \mathbf{1}_{[-1,1]}(x^1) \mathbf{1}_{[-1,1]}(x^2)$
 - Gaussian component: $\pi_2 = 0.5$, $\mu_2 = (3.3, 0)'$, $\Sigma_2 = I$
- 50 simulated data sets of size $n = 200$



| K | 1 | 2 | 3 | 4 | 5 |
|-----|---|------------|---|----|---|
| BIC | . | 60 | . | 32 | 8 |
| ICL | . | 100 | . | . | . |

Outline

1 Mixture models as a multi-purpose tool

2 Gaussian case

3 Mixed data case

4 Ranking data case

5 RMixtComp in practice

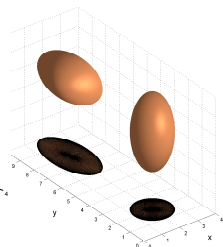
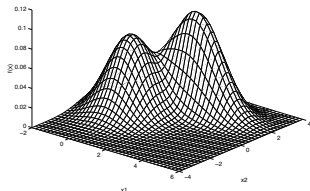
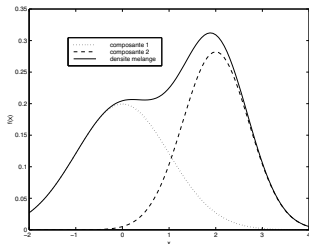
6 Rankcluster in practice

Gaussian mixtures are classical
for numerical data ($\mathcal{X} = \mathbb{R}^d$)

d -variate Gaussian mixture model

$$p(\mathbf{x}_1; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \underbrace{\frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_k)\right)}_{p(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

$$p(\cdot; \boldsymbol{\alpha}_k) = N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{where} \quad \boldsymbol{\alpha}_k = \left(\underbrace{\boldsymbol{\mu}_k}_{\text{center}}, \underbrace{\boldsymbol{\Sigma}_k}_{\text{dispersion}} \right)$$



EM with complete data \mathbf{x} : E-step

Just compute all conditional probabilities $t_{ik}(\boldsymbol{\theta}^{(q)}) = t_k(\mathbf{x}_i; \boldsymbol{\theta}^{(q)})$.

EM with complete data \mathbf{x} : M-step

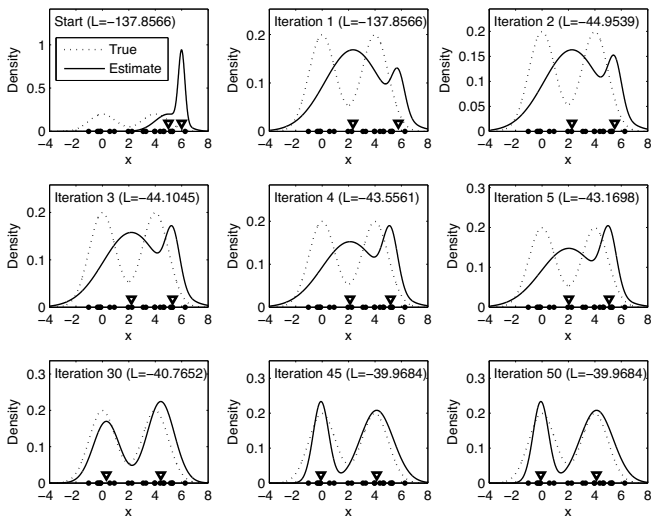
$$n_k^{(q)} = \sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(q)})$$

$$\pi_k^{(q+1)} = \frac{n_k^{(q)}}{n}$$

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(q)}) \mathbf{x}_i \right)$$

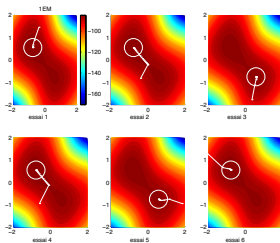
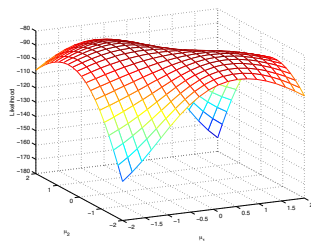
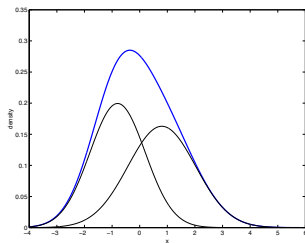
$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^{(q)}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' \right)$$

Example of EM in the univariate case with complete data x



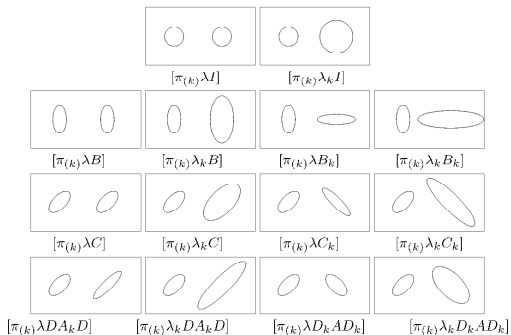
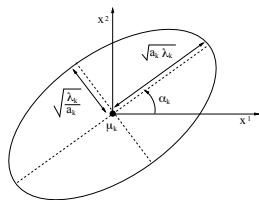
Note : low at the beginning but increase of the log-likelihood

Local maxima



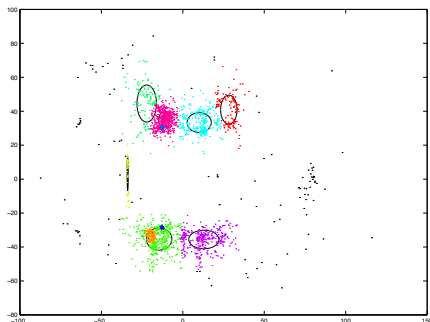
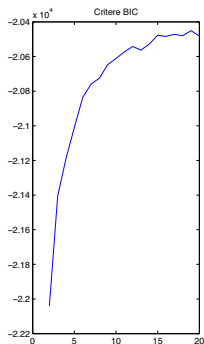
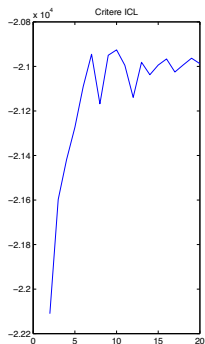
Constraints on Σ_k

$$\Sigma_k = \underbrace{\lambda_k}_{\text{volume}} \cdot \underbrace{\mathbf{D}_k}_{\text{orientation}} \cdot \underbrace{\mathbf{A}_k}_{\text{shape}} \cdot \mathbf{D}'_k$$



Acoustic emission control example with diagonal constraints

- **Data:** $n = 2\,061$ event locations in a rectangle of \mathbb{R}^2 representing the vessel
- **Model:** Diagonal Gaussian mixture + uniform (noise)
- **Groups:** sound locations = vessel defects



EM with incomplete data \mathbf{x}^o : E-step

θ and θ^+ the parameters for two successive steps (*idem* for missing data)

$$z_{ik}^+ = p(z_{ik} = 1 | \mathbf{x}_i^o; \theta) = \frac{\pi_k p(\mathbf{x}_i^o; \mu_k, \Sigma_k)}{p(\mathbf{x}_i^o; \theta)}$$

$$\mathbf{x}_{ik}^{m+} = E[\mathbf{x}_i^m | \mathbf{x}_i^o, z_{ik} = 1; \theta] = \mu_{ik}^m + \Sigma_{ik}^{mo} (\Sigma_{ik}^{oo})^{-1} (\mathbf{x}_i^o - \mu_{ik}^o)$$

where

- $o_i \subseteq \{1, \dots, d\}$ the set of the observed variables for \mathbf{x}_i
- \mathbf{x}_i^o the corresponding observed data
- m_i the set of the missing variables for \mathbf{x}_i
- μ_{ik}^o the sub-vector of μ_k associated to index o_i (the same for m_i)
- Σ_{ik}^o the sub-matrix of Σ_k associated to row o_i and columns m_i (the same for any other combination)

Interpretation

- z_{ik}^+ : conditional probability membership given the available information \mathbf{x}_i^o .
- \mathbf{x}_{ik}^{m+} : conditional imputation of the missing data given the cluster.

EM with incomplete data \mathbf{x}^o : M-step

$$\begin{aligned}\pi_k^+ &= \frac{1}{n_k^+} \sum_{i=1}^n z_{ik}^+, \quad \boldsymbol{\mu}_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n z_{ik}^+ \mathbf{x}_{ik}^+ \\ \boldsymbol{\Sigma}_k^+ &= \frac{1}{n_k^+} \sum_{i=1}^n z_{ik}^+ \left[(\mathbf{x}_{ik}^+ - \boldsymbol{\mu}_k^+) (\mathbf{x}_{ik}^+ - \boldsymbol{\mu}_k^+)' + \boldsymbol{\Sigma}_{ik}^+ \right]\end{aligned}$$

where $n_k^+ = \sum_{i=1}^n z_{ik}^+$, $\mathbf{x}_{ik}^+ = \begin{pmatrix} \mathbf{x}_i^o \\ \mathbf{x}_{ik}^{m+} \end{pmatrix}$, $\boldsymbol{\Sigma}_{ik}^+ = \begin{pmatrix} \mathbf{0}_i^o & \mathbf{0}_i^{om} \\ \mathbf{0}_i^{mo} & \boldsymbol{\Sigma}_{ik}^{m+} \end{pmatrix}$ with $\mathbf{0}$ the $d \times d$ null matrix, and $\boldsymbol{\Sigma}_{ik}^{m+} = \boldsymbol{\Sigma}_{ik}^{mo} (\boldsymbol{\Sigma}_{ik}^o)^{-1} \boldsymbol{\Sigma}_{ik}^{om}$.

Interpretation of $\boldsymbol{\Sigma}_{ik}^{m+}$

Variance correction due to the under-estimation of variability caused by the imputation of missing data.

Outline

1 Mixture models as a multi-purpose tool

2 Gaussian case

3 Mixed data case

4 Ranking data case

5 RMixtComp in practice

6 Rankcluster in practice

Categorical data: latent class model

- **categorical variables:** d variables with l_j modalities each, $\mathbf{x}_{ij} \in \{0, 1\}^{l_j}$ and

$$x_{jh} = 1 \Leftrightarrow \text{variable } j \text{ of } \mathbf{x}_i \text{ takes modality } h$$

- **Conditional independence:**

$$p(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^d \prod_{h=1}^{l_j} (\alpha_{kjh})^{x_{ijh}}$$

and

$$\alpha_{kjh} = p(x_{ijh} = 1 | z_{ik} = 1)$$

with $\boldsymbol{\alpha}_k = (\alpha_{kjh}; j = 1, \dots, d; h = 1, \dots, l_j)$

EM illustration with binary data: SPAM E-mail Database⁴

- $n = 4\,601$ e-mails composed by 1813 “spams” and 2 788 “good e-mails”
- $d = 48 + 6 = 54$ continuous descriptors³
 - 48 percentages that a given **word** appears in an e-mail (“make”, “you’...”)
 - 6 percentages that a given **char** appears in an e-mail (“;”, “\$”...)
- Transformation of continuous descriptors into **binary descriptors**

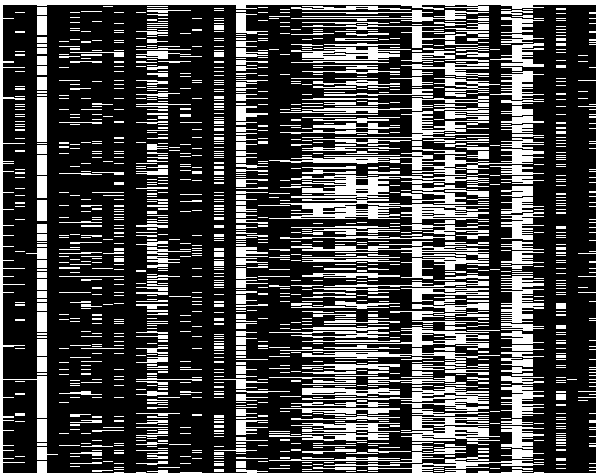
$$x_{ij} = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

³There are 3 other continuous descriptors we do not use

⁴<https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/>

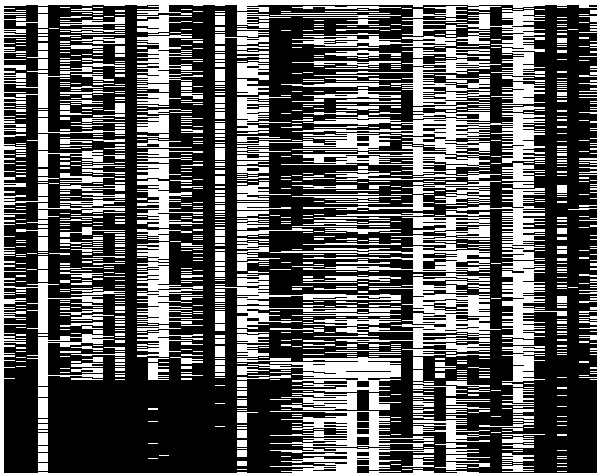
An EM run with a binary data set

Initial binary data



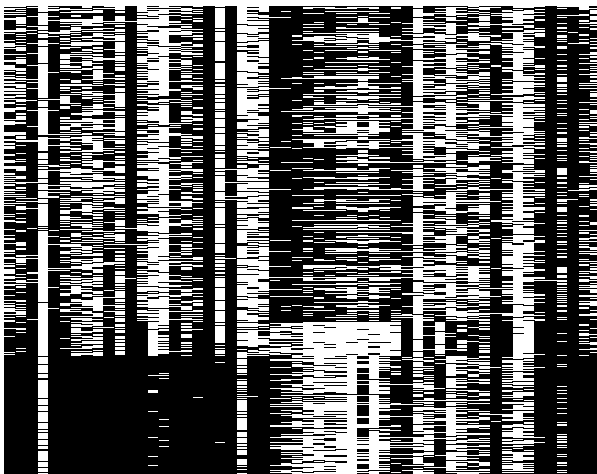
An EM run with a binary data set

Iteration 1



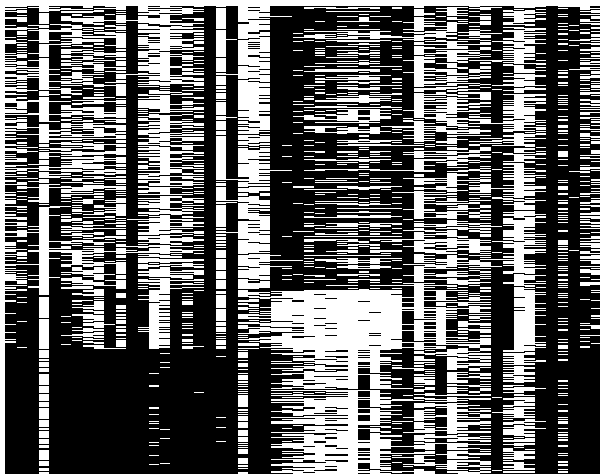
An EM run with a binary data set

Iteration 2



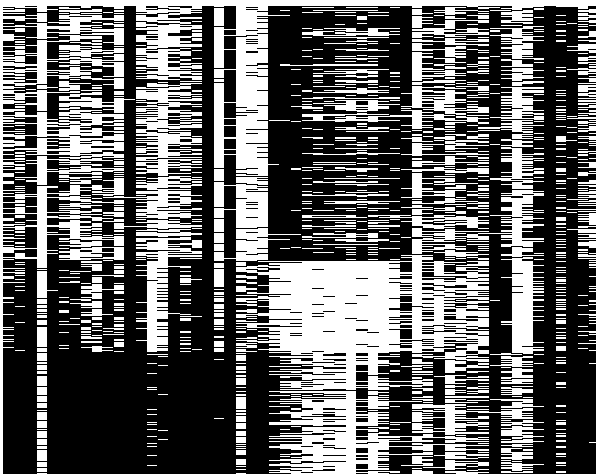
An EM run with a binary data set

Iteration 3



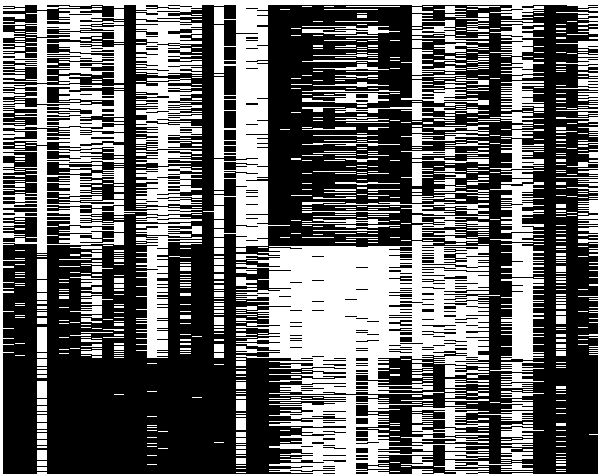
An EM run with a binary data set

Iteration 4



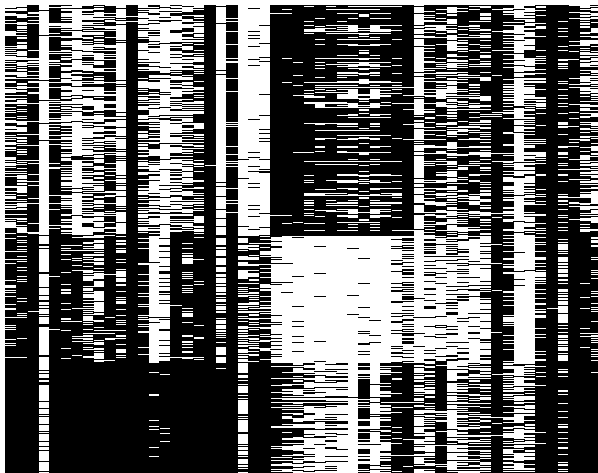
An EM run with a binary data set

Iteration 5



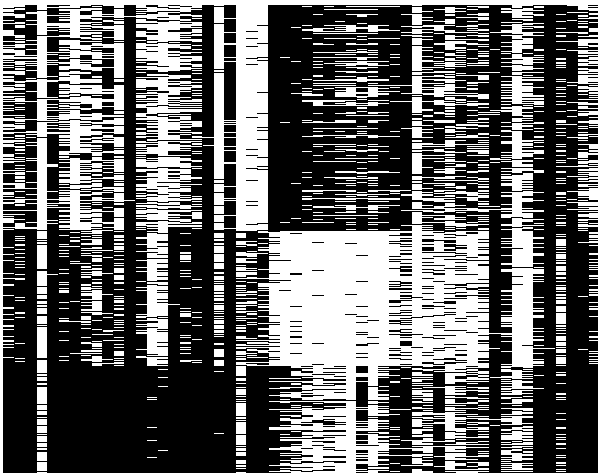
An EM run with a binary data set

Iteration 6



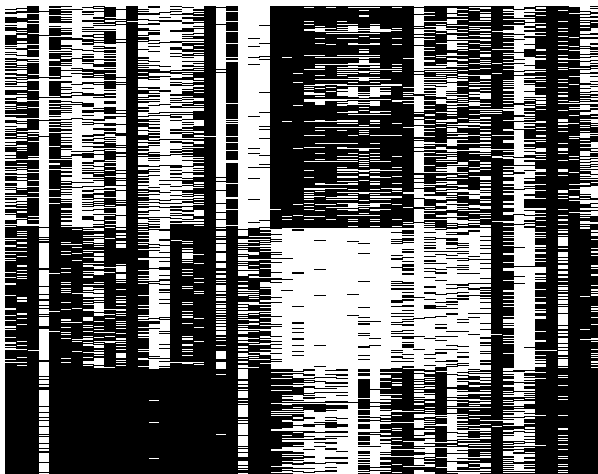
An EM run with a binary data set

Iteration 7



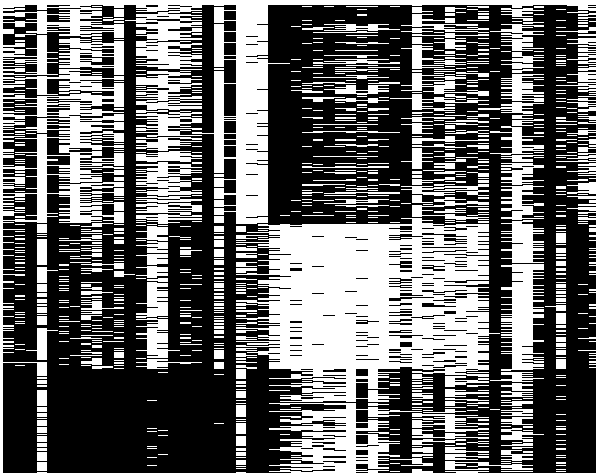
An EM run with a binary data set

Iteration 8



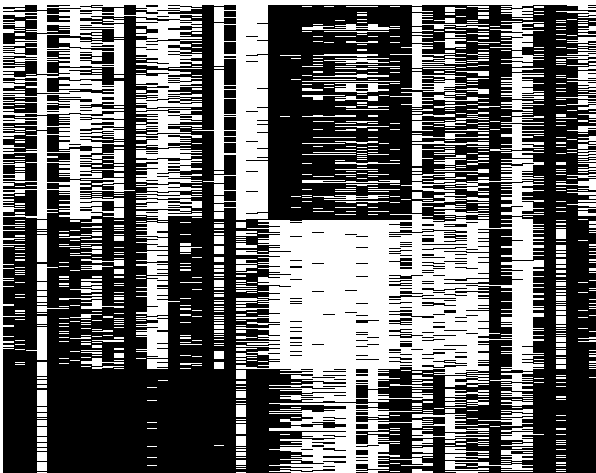
An EM run with a binary data set

Iteration 9



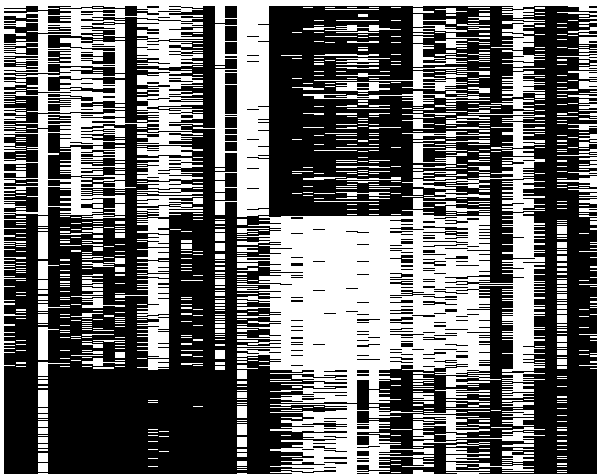
An EM run with a binary data set

Iteration 10



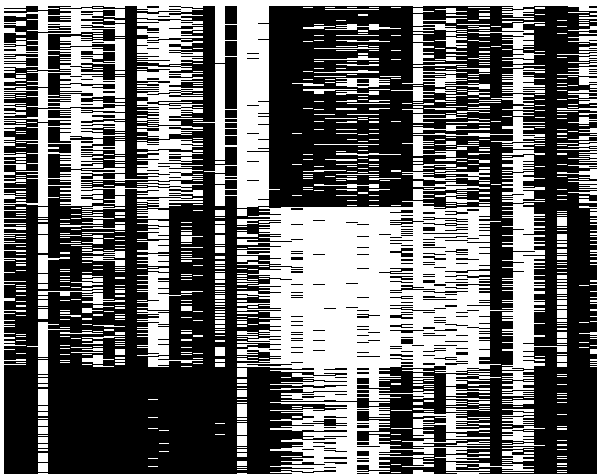
An EM run with a binary data set

Iteration 11



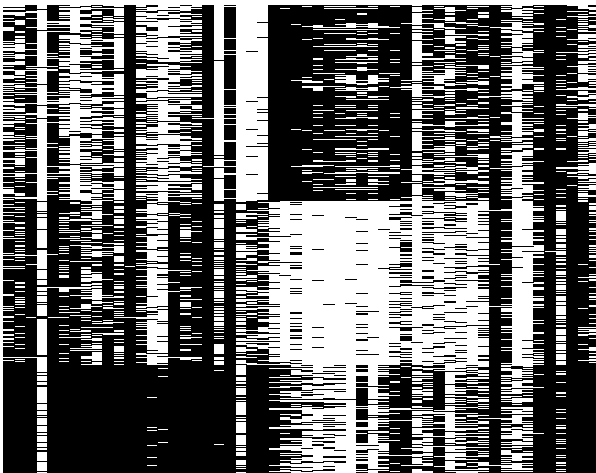
An EM run with a binary data set

Iteration 12



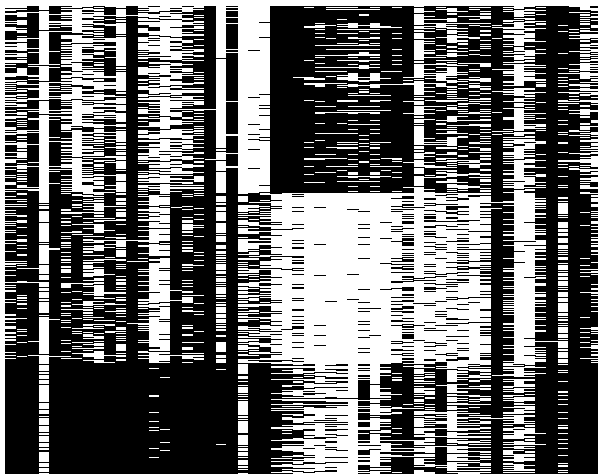
An EM run with a binary data set

Iteration 13



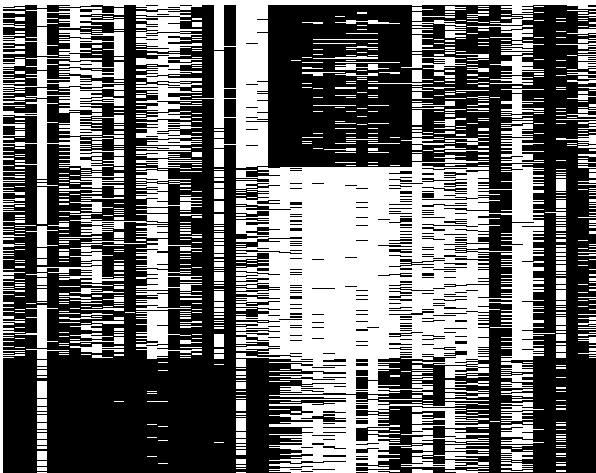
An EM run with a binary data set

Iteration 14



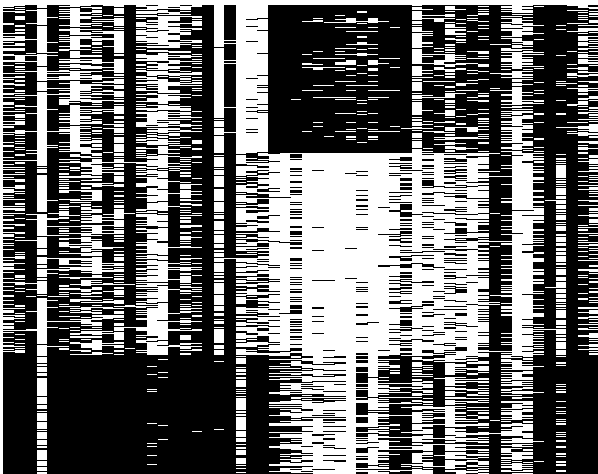
An EM run with a binary data set

Iteration 15



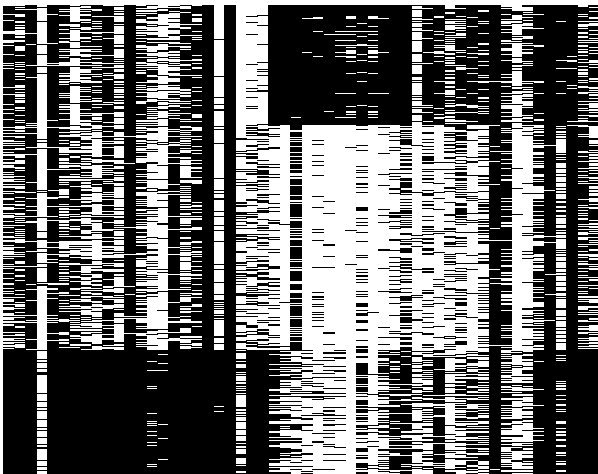
An EM run with a binary data set

Iteration 16



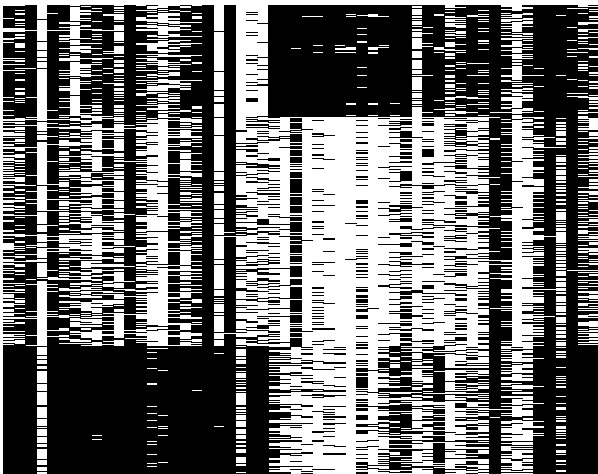
An EM run with a binary data set

Iteration 17



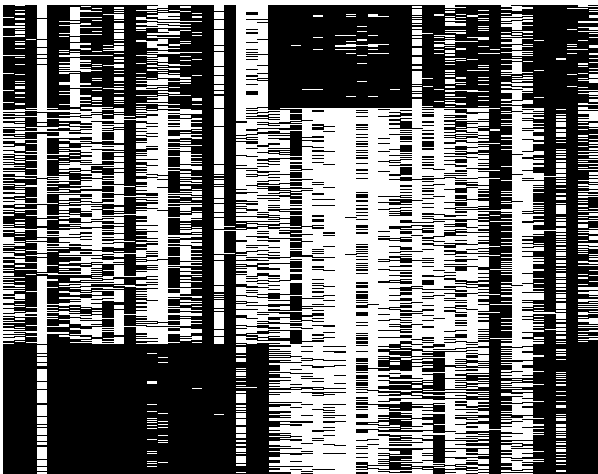
An EM run with a binary data set

Iteration 18



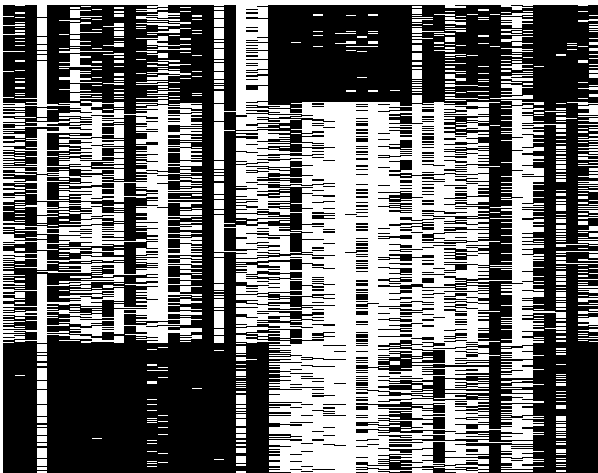
An EM run with a binary data set

Iteration 19

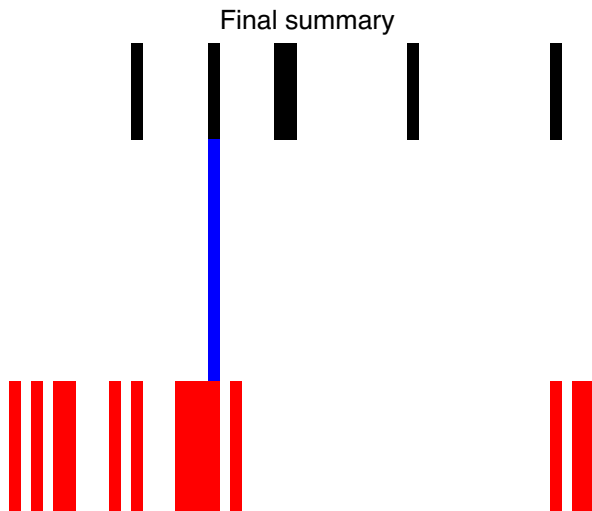


An EM run with a binary data set

Iteration 20



An EM run with a binary data set



Integer: Poisson mixture model

- **integer variables:** d variables $\mathbf{x}_{ij} \in \mathbb{N}$
- **Intra conditional independence:**

$$p(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^d \frac{(\alpha_{kj})^{x_{ij}}}{\alpha_{kj}!} e^{-\alpha_{kj}}$$

Mixed data: classical approaches

Usually, unify data type by transformation:

- Quantify continuous variables: [loose some information](#)
- MCA of categorical variable: [loose the meaning](#)
- ...

Proposal

Model-based directly on [raw data](#)

Mixed data: conditional independence everywhere

The aim is to combine continuous, categorical and integer data

$$\mathbf{x}_1 = (\mathbf{x}_1^{cont}, \mathbf{x}_1^{cat}, \mathbf{x}_1^{int})$$

The proposed solution is to mixed all types by **inter-type conditional independence**

$$p(\mathbf{x}_1; \alpha_k) = p(\mathbf{x}_1^{cont}; \alpha_k^{cont}) \times p(\mathbf{x}_1^{cat}; \alpha_k^{cat}) \times p(\mathbf{x}_1^{int}; \alpha_k^{int})$$

In addition, for symmetry between types, **intra-type conditional independence**

Only need to define the univariate pdf for each variable type!

- **Continuous**: Gaussian
- **Categorical**: multinomial
- **Integer**: Poisson

Example of mixed case: Cancer dataset with more missing data

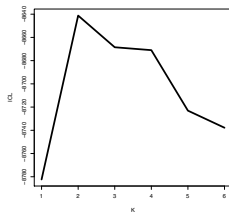
Add artificially $\approx 30\%$ missing data with a MCAR design

Then compare two strategies of imputation:

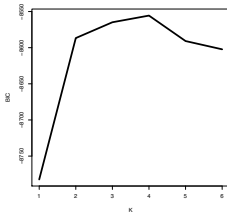
- Strategy “mice”: dataset completed by mice

```
> data.imp=mice(data)
> data.comp.mice=complete(data.imp)
```

- Strategy “full MixtComp”: MixtComp on the observed (no completed) dataset



ICL
 $\hat{K} = 2$



BIC
 $\hat{K} = 4$

Example of mixed case: imputation accuracy

- **Continuous variables:** mean of absolute difference between x and \hat{x}

| var. | mice | MixtComp ($K = 2$) | MixtComp ($K = 4$) |
|-------------|----------|----------------------|----------------------|
| Age | 8.907143 | 5.546571 | 5.526861 |
| Wt | 13.51656 | 9.779485 | 9.731182 |
| SBP | 2.103226 | 1.788152 | 1.795820 |
| DBP | 1.317568 | 1.165201 | 1.169672 |
| HG | 21.67568 | 14.83514 | 14.51291 |
| SZ | 1.714899 | 1.160546 | 1.158105 |
| SG | 1.979866 | 1.386841 | 1.416053 |
| AP | 1.359299 | 1.027513 | 1.009126 |
| Global mean | 6.5718 | 4.5862 | 4.5400 |

- **Categorical variable:** mean of the proportion of difference between x and \hat{x}

| var. | mice | MixtComp ($K = 2$) | MixtComp ($K = 4$) |
|-------------|-----------|----------------------|----------------------|
| PF | 0.1904762 | 0.0952381 | 0.0952381 |
| HX | 0.4121622 | 0.4391892 | 0.4121622 |
| EKG | 0.7564103 | 0.6858974 | 0.7179487 |
| BM | 0.1081081 | 0.1486486 | 0.1216216 |
| Global mean | 0.3668 | 0.3422 | 0.3367 |

Outline

1 Mixture models as a multi-purpose tool

2 Gaussian case

3 Mixed data case

4 Ranking data case

5 RMixtComp in practice

6 Rankcluster in practice

Notations *ranking and ordering*

Rank definition

A rank consists of sorting l objects following a preference order.

Example: sort three holidays places

\mathcal{O}_1 : countryside, \mathcal{O}_2 : mountain et \mathcal{O}_3 : sea

\Rightarrow 1st sea, 2nd countryside et 3rd mountain

Notations

- **Ordering:** $\mathbf{x} = (3, 1, 2) = (\overset{1^{\text{st}}}{\mathcal{O}_3}, \overset{2^{\text{nd}}}{\mathcal{O}_1}, \overset{3^{\text{th}}}{\mathcal{O}_2})$
- **Ranking:** $\mathbf{x}^{-1} = (2, 3, 1) = (\overset{\mathcal{O}_1}{2^{\text{nd}}}, \overset{\mathcal{O}_2}{3^{\text{th}}}, \overset{\mathcal{O}_3}{1^{\text{st}}})$

$\mathbf{x}, \mathbf{x}^{-1} \in \mathcal{P}_l$ (permutations of the first l integers).

Interest of ranking data

Human activities implying preferences or choices

Web pages sorting

Sociology

Economy

Biology

Sport

Politics

Psychology

Marketing ...

Missing data are frequent

- When many objects to be sorted, some rankings can be incomplete
- Ex-æquo values can also be seen as incomplete rankings

Univariate ISR (Insertion Sorting Rank) model

Données

- $\mathbf{x} = (x_1, \dots, x_I)$: observed rankings
- $\mathbf{y} = (y_1, \dots, y_I)$: latent order presentation of objects

Hypothèse

\mathbf{x} arises from a **insertion sorting algorithm** with parameters

- $\mu = (\mu_1, \dots, \mu_I)$: reference ranking (**position**)
- $\alpha \in [0, 1]$: proba. of good comparison by pair (**dispersion**)

Associated distribution, after averaging over unknown y

$$\text{pr}(\mathbf{x}; \mu, \pi) = \frac{1}{I} \sum_y \alpha^{\text{good}(\mathbf{x}, \mathbf{y}, \mu)} (1 - \alpha)^{\text{bad}(\mathbf{x}, \mathbf{y}, \mu)}$$

ISR properties

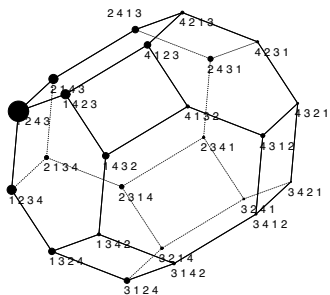
Symetry

$$\text{pr}(\mathbf{x}; \bar{\mu}, 1 - \alpha) = \text{pr}(\mathbf{x}; \mu, \alpha) \quad \Rightarrow \quad \alpha \in \left[\frac{1}{2}, 1\right]$$

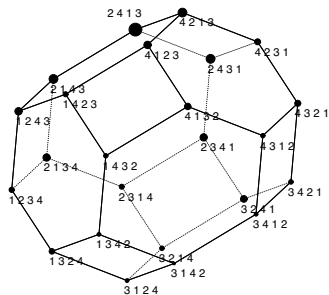
Bon comportement

- μ is the mode and $\bar{\mu}$ is the anti-mode ($\alpha > \frac{1}{2}$)
- $\text{pr}(\mu; \mu, \alpha) - \text{pr}(\mathbf{x}; \mu, \alpha)$ is a **non-decreasing function of α** :
the largest is α , the sharper is the distribution around its mode
- the parameters (μ, α) are **identifiable** if $\alpha > \frac{1}{2}$
- the distribution is **uniform** for $\pi = \frac{1}{2}$, **Dirac** at μ if $\alpha = 1$

ISR distribution illustration



$\mu = (1, 2, 4, 3)$ et $\alpha = 0.83$



$\mu = (2, 4, 1, 3)$ et $\alpha = 0.68$

Extension to the ISR multivariate mixture model⁵

Multivariate ranks

- Dimension d : $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$
- l_j objects per dimension ($1 \leq j \leq d$): $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijl_j})$

K -mixture of multivariate ISR

Hyp. of **class conditional independence**

$$p(\mathbf{x}_i; \boldsymbol{\theta}) = \underbrace{\sum_{k=1}^K \pi_k}_{\text{mixture}} \underbrace{\prod_{j=1}^d \overbrace{p(\mathbf{x}_{ij}; \mu_{kj}, \alpha_{kj})}^{\text{ISR}(\mu_{kj}, \alpha_{kj}) \text{ univariate}}}_{\text{ISR multivariate}}$$

- Proportions π_k : $\pi_k \in [0, 1]$ et $\sum_{k=1}^K \pi_k = 1$
- Whole parameter: $\boldsymbol{\theta} = (\alpha_{kj}, \mu_{kj}, \pi_k)_{k=1, \dots, K; j=1, \dots, d}$

⁵J. Jacques and C. Biernacki (2014). Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference*, 149, 201–217.

Maximum likelihood estimation

- EM & SEM: impossible since combinatorics difficulty because of permutations
- SEM within Gibbs: embed a Gibbs algorithm within the SE-step

Utility of ISR mixtures: data

4 quizz proposed to 30 students (GIS4 - Polytech'Lille)

- Sort chronologically these writers

1. Hugo 2. Molière 3. Camus 4. Rousseau

- Sort these countries by increasing order of their number of victories at the football world cup

1. France 2. Germany 3. Brasil 4. Italy

- Sort these numbers by increasing order

1. $\pi/3$ 2. $\ln 1$ 3. e^2 4. $(1 + \sqrt{5})/2$

- Sort chronologically these movies of Tarantino

1. Inglorious Basterds 2. Pulp Fiction 3. Reservoir Dogs 4. Jackie Brown

Utility of ISR mixtures: Quizz 1

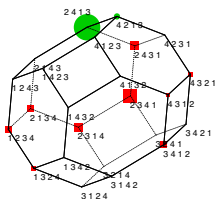
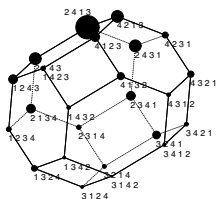
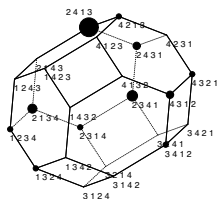
Sort chronologically these writers

1. Hugo

2. Molière

3. Camus

4. Rousseau

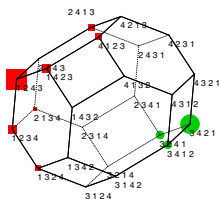
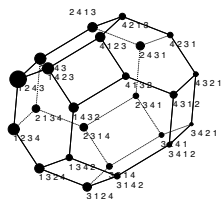
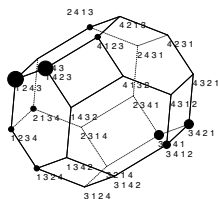


Empiric

| | ISR | Mixture of ISR | |
|------------|--------------|----------------|--------------|
| μ_k | (2, 4, 1, 3) | (2, 4, 1, 3) | (2, 3, 4, 1) |
| α_k | 0.80 | 0.95 | 0.71 |
| π_k | | 0.5 | 0.5 |
| BIC | 152.3 | 148.0 | |

Utility of ISR mixtures: Quizz 2

Sort these countries by increasing order of their number of victories at the football world cup : 1. France, 2. Germany, 3. Brasil, 4. Italy



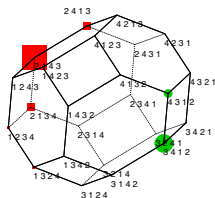
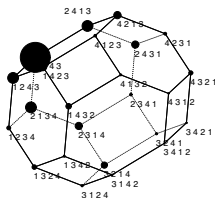
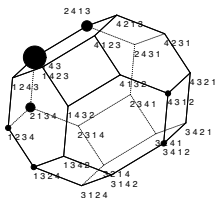
Empiric

| | ISR | Mixture of ISR | |
|------------|--------------|----------------|--------------|
| μ_k | (1, 2, 4, 3) | (1, 2, 4, 3) | (3, 4, 2, 1) |
| α_k | 0.69 | 0.85 | 0.84 |
| π_k | | 0.73 | 0.27 |
| BIC | 179.1 | 160.6 | |

Utility of ISR mixtures: Quizz 3

Sort these numbers by increasing order

1. $\pi/3$ 2. $\ln 1$ 3. e^2 4. $(1 + \sqrt{5})/2$



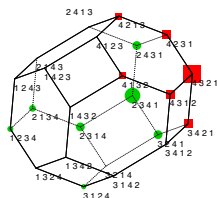
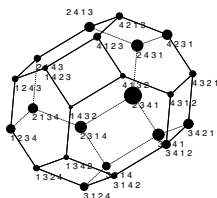
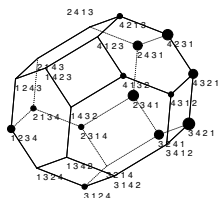
Empiric

| | ISR | Mixture of ISR | |
|------------|--------------|----------------|--------------|
| μ_k | (2, 1, 4, 3) | (2, 1, 4, 3) | (3, 4, 1, 2) |
| α_k | 0.69 | 0.92 | 0.82 |
| π_k | | 0.92 | 0.08 |
| BIC | 111.4 | 106.3 | |

Utility of ISR mixtures: Quizz 4

Sort chronologically these movies of Tarantino

1.Inglorious Bastards 2.Pulp Fiction 3.Reservoir Dogs 4.Jackie Brown



Empiric

| | ISR | Mixture of ISR | |
|------------|--------------|----------------|--------------|
| μ_k | (2, 3, 4, 1) | (2, 3, 4, 1) | (4, 3, 2, 1) |
| α_k | 0.69 | 0.76 | 0.79 |
| π_k | | 0.61 | 0.39 |
| BIC | 177.0 | 173.8 | |

Social comparison theory (1/4)

■ Questionnaire

Which things do you prefer to compare with other children of your age? Put a 1 in front of what you prefer to compare most, a 2 in front of what you prefer next, and so on. **More than 3 is not necessary, but is allowed:**

- | | |
|---------------------------------|---|
| 1 your popularity, | 8 your grades at school, |
| 2 how well you do in sports, | 9 how well you can express your opinions, |
| 3 your appearance, | 10 your hobby's, |
| 4 how much money you can spend, | 11 how "courageous" you are, |
| 5 how you are feeling, | 12 how smart you are, |
| 6 your parents, | 13 the kind of friends you have. |
| 7 your clothes, | |

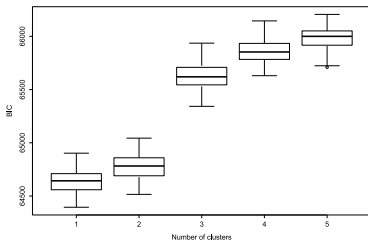
■ Data

- $n = 1\,567$ students
- $l = 13$ objects to be compared ($d = 1$)
- 85% of partial ranks
- among 15% of full ranks, 20% contain ties

Social comparison theory (2/4)

Package Rankcluster

```
R> res=rankclust(x,1,1:5)
```



- Log-likelihood estimated by the harmonic mean (intractable closed-form)
- BIC “hesitates” between **one and two groups**: many missing data...

Social comparison theory (3/4)

```
*****  
Number of clusters: 1  
*****  
proportion: 1  
mu:  
c11  8 12 1 11 6 13 3 2 4 7 9 10 5  
alpha:  
c11 0.6478778
```

- π low $\Rightarrow \mu$ not significant \Rightarrow nearly uniform distribution
- Example 1st/2nd: $P(\mathcal{O}_3 \succ \mathcal{O}_8) = 0.54 \rightarrow$ low!
- Example 1st/last: $P(\mathcal{O}_3 \succ \mathcal{O}_6) = 0.62 \rightarrow$ low!
- Conclusion: no **global** preference

Social comparison theory (4/4)

```

*****
Number of clusters: 2
*****
proportion: 0.9261865 0.0738135
mu:
c11    9 5 3 10  4 8 6 1 11  7 12 13 2
c12    4 1 2  8 13 7 5 3  9 12 11 10 6
alpha:
c11 0.6583201
c12 0.7977076

```

- Group 1 large ($\pi_1 = 93\%$) :
 - Looks like the one-group case
 - No real preference ($\alpha_1 = 0.65$ low)
- Group 2 small ($\pi_1 = 7\%$) :
 - Preference more apparent ($\alpha_2 = 0.80$ greater)
 - Example 1st/2nd: $P(\mathcal{O}_2 \succ \mathcal{O}_3) = 0.64 \rightarrow$ more significant!
 - Example 1st/last: $P(\mathcal{O}_3 \succ \mathcal{O}_5) = 0.79 \rightarrow$ significant !
- Conclusion: no **global** preference, but **locally** a small group

sports \succ appearance \succ grades \succ popularity \succ ...

Eurovision Song Contest (1/2)

Principle

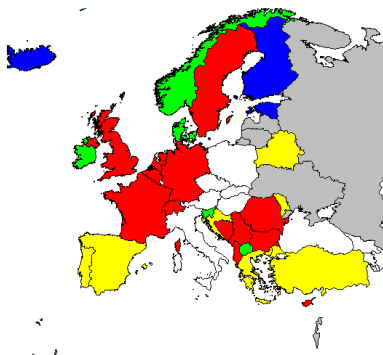
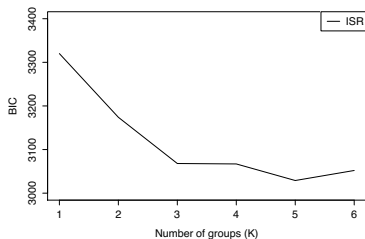
- Biggest musical competition in the world (around 40 countries)
- Each member country submits a song broadcast live...
- ... then rank these 10 favorite foreign songs

Data

- Votes of $n = 34$ participating countries between 2007 and 2012 ($d = 6$ years)
- Only 8 countries participated in the 6 finals:
1: France, 2: Germany, 3: Greece, 4: Romania, 5: Russia, 6: Spain, 7: Ukraine, 8: UK
- Only the votes for these 8 countries are considered: $l_j = 8$
- 57.7% of ranking elements are missing!

Eurovision Song Contest (2/2)

Package Rankclust en C++ interfaced with R



Revelation of geographic alliances...

Outline

1 Mixture models as a multi-purpose tool

2 Gaussian case

3 Mixed data case

4 Ranking data case

5 RMixtComp in practice

6 Rankcluster in practice

RMixtComp: pitch

- Mixture Models with Mixed and (Partially) Missing Data
- 8 models for real, categorical, counting, functional and ranking data
- SEM algorithm with MAR assumption
- Available on the CRAN repository:
<https://cran.r-project.org/web/packages/RMixtComp/index.html>

RMixtComp: syntax/allowed missing data

allowed missing value types for each model

| | Categorical_pjk | Gaussian_sjk | Poisson_k | LatentClass |
|--|-----------------|--------------|-----------|-------------|
| ? (completely missing) | X | X | X | X |
| $\{a, b, c\}$ (finite number of values authorized) | X | | | X |
| $[a : b]$ (bounded interval) | | X | | |
| $[-inf : b]$ (semi-bounded interval) | | X | | |
| $[a : +inf]$ (semi-bounded interval) | | X | | |

RMixtComp: overview of the output R format

```

res
  strategy
    nbTrialInInit
    nbBurnInIter
    nbIter
    nbGibbsBurnInIter
    nbGibbsIter
  mixture
    nbCluster
    nbFreeParameters
    lnObservedLikelihood
    lnSemiCompletedLikelihood
    lnCompletedLikelihood
    BIC
    ICL
    runTime
    nbSample
    warnLog
  variable
    data
      z_class
        completed !!! <- imputed classes
        stat !!! <- a posteriori distribution of class for each individual (= p(z_i / x_i))
        categorical1
          completed
          stat
        categorical2, etc ...
    param
      z_class
        stat !!! <- model proportions and quantiles
        log
        categorical1
          stat
          log
        categorical2, etc ...

```

Note that the `z_class` variable contains all the information pertaining to the latent classes:

- `res$variable$data$sample$completed` contains the imputation for the class, \hat{z}_i
- `res$variable$data$sample$stat` contains the estimated posteriori probabilities, \hat{t}_{ik}
- `res$variable$paramz_classstat` contains the proportions, $\hat{\pi}_k$

Prostate data set: overview of data file prostate.csv

data.csv - Op

Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

Arial 10 G I S

A1 z_class

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|----|---------|-----|-----|----|----|-----|-----|-----|-----|--------|----|----------|----|---|
| 1 | z_class | Age | Wt | PF | HX | SBP | DBP | EKG | HG | SZ | SG | AP | BM | |
| 2 | ? | 75 | 76 | 1 | 1 | 15 | 9 | 5 | 138 | 1.4142 | 8 | 1.0986 | 1 | |
| 3 | ? | 54 | 116 | 1 | 1 | 13 | 7 | 4 | 146 | 6.4807 | ? | 1.9459 | 1 | |
| 4 | ? | 69 | 102 | 1 | 2 | 14 | 8 | 5 | 134 | 1.7321 | 9 | 1.0986 | 1 | |
| 5 | ? | 75 | 94 | 2 | 2 | 14 | 7 | 2 | 176 | | 2 | 82.1972 | 1 | |
| 6 | ? | 67 | 99 | 1 | 1 | 17 | 10 | 1 | 134 | 5.831 | 8 | 1.6094 | 1 | |
| 7 | ? | 71 | 98 | 1 | 1 | 19 | 10 | 1 | 151 | 3.1623 | 11 | 1.7918 | 1 | |
| 8 | ? | 75 | 100 | 1 | 1 | 14 | 10 | 2 | 130 | 3.6056 | 9 | 2.0794 | 1 | |
| 9 | ? | 73 | 114 | 1 | 2 | 17 | 11 | 5 | 126 | 1.7321 | 9 | 1.7918 | 1 | |
| 10 | ? | 60 | 110 | 1 | 1 | 12 | 8 | 1 | 146 | | 2 | 10.19459 | 1 | |
| 11 | ? | 78 | 107 | 1 | 2 | 13 | 8 | 6 | 130 | 4.5826 | 6 | 1.3863 | 1 | |
| 12 | ? | 77 | 89 | 1 | 1 | 15 | 8 | 1 | 156 | 1.7321 | 8 | 1.7918 | 1 | |
| 13 | ? | 74 | 105 | 1 | 2 | 18 | 14 | 1 | 136 | 2.4495 | 8 | 1.3863 | 1 | |
| 14 | ? | 74 | 107 | 1 | 1 | 14 | 9 | 6 | 144 | 2.4495 | 9 | 1.0986 | 1 | |
| 15 | ? | 55 | 112 | 1 | 2 | 16 | 9 | 5 | 139 | | 2 | 92.3026 | 1 | |
| 16 | ? | 73 | 88 | 1 | 1 | 19 | 10 | 5 | 120 | 3.873 | 10 | 1.7918 | 1 | |
| 17 | ? | 87 | 81 | 2 | 2 | 17 | 12 | 3 | 134 | 1.7321 | 9 | 1.3863 | 1 | |
| 18 | ? | 64 | 90 | 1 | 1 | 14 | 8 | 1 | 162 | 2.4495 | 9 | 1.9459 | 1 | |
| 19 | ? | 79 | 104 | 1 | 1 | 13 | 8 | 2 | 150 | 2.2361 | 8 | 1.6094 | 1 | |
| 20 | ? | 62 | 90 | 1 | 2 | 13 | 8 | 2 | 144 | 1.4142 | 9 | 1.9459 | 1 | |
| 21 | ? | 74 | 107 | 1 | 1 | 14 | 9 | 6 | 144 | 2.4495 | 9 | 1.0986 | 1 | |

Prostate data set: overview of the variable descriptors

prostate_descriptor.csv - OpenOffice.org Calc

Echier Édition Affichage Insertion Format Outils Données Fenêtre Aide

Arial 10 G / S

A1 z_class

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|-------------|--------------|--------------|-----------------|-----------------|--------------|--------------|-----------------|--------------|--------------|--------------|--------------|----------|
| 1 | z_class | Age | Wt | PF | HX | SBP | DBP | EKG | HG | SZ | SG | AP | BM |
| 2 | LatentClass | Gaussian_sjk | Gaussian_sjk | Categorical_pjk | Categorical_pjk | Gaussian_sjk | Gaussian_sjk | Categorical_pjk | Gaussian_sjk | Gaussian_sjk | Gaussian_sjk | Gaussian_sjk | Categori |
| 3 | | | | | | | | | | | | | |

Prostate data set: overview of the output multiple imputation by MixtComp

```

> res$variable$data$EKGS$completed[[5]]
[1] 1
> res$variable$data$EKGS$stat[[1]]
[[1]]
[1] 5

[[2]]
[1] 1

[[3]]
[1] 0.41

[[4]]
[1] 5

[[5]]
[1] 0.29

[[6]]
[1] 6

> res$variable$data$Age$completed[8]
[1] 70.62032
> res$variable$data$Age$stat[[1]]
[[1]]
[1] 8

[[2]]
[1] 70.62032

[[3]]
[1] 58.24255

[[4]]
[1] 83.86463

[[7]]
[1] 0.15

[[8]]
[1] 2

[[9]]
[1] 0.07

[[10]]
[1] 3

[[11]]
[1] 0.05

```

cont.

cat.

Prostate data set: ready for the practical activity with RMixtComp!

■ Learning step:

- Run RMixtComp with $K \in \{1, \dots, 8\}$
- Check graphically K value retained by BIC and ICL
- Check the imputed values (and there confidence interval) of continuous/categorical missing values

■ Prediction step:

- Transform Patient 1 as follows: Age is missing, Wt is uncertain within [70,80], EKG is uncertain within {4, 5}
- For this transformed Patient 1: retain $K = 3$, estimate his/her class, Age, Wt, EKG (and have a look at the related confidence intervals)

Prostate data set: R code for learning

```

data <- read.table("prostate.csv", sep = ";", header = TRUE)
head(data)
library(RMixtComp)

# Define the distribution used for each variable.
model <- list(Age = "Gaussian", Wt = "Gaussian", PF = "Multinomial",
             HX = "Multinomial", SEP = "Gaussian", DBP = "Gaussian",
             EKG = "Multinomial", HG = "Gaussian", SZ = "Gaussian",
             SG = "Gaussian", AP = "Gaussian", BM = "Multinomial")

# Define the SEM algorithm's parameters
algo <- list(nbBurnInIter = 50,
            nbIter = 100,
            nbGibbsBurnInIter = 50,
            nbGibbsIter = 100,
            nInitPerClass = floor(nrow(data)/2),
            nSemTry = 5,
            confidenceLevel = 0.95,
            ratioStableCriterion = 0.99,
            nStableCriterion = 10)

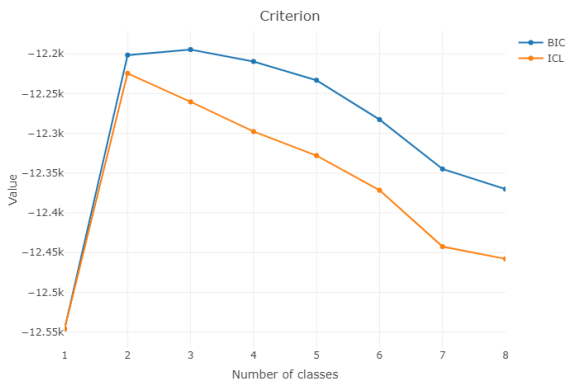
# Choose the desired number of classes and the number of runs for each given number of classes.
nClass <- 1:8
nRun <- 3
res <- mixtCompLearn(data, model, algo, nClass = nClass, criterion = "ICL", nRun = nRun, nCore = 1)

# Draw the criterion value (BIC and ICL) for each model that was built. The higher the value (close to 0) the better the model.
plotCrit(res, pkg = "plotly")

# See estimation of all the missing Age values (idem for other variables)
res$variable$data$Age$completed # imputed
res$variable$data$Age$stat # confidence interval
res$variable$data$BM$completed # imputed
res$variable$data$BM$stat # confidence interval

```

Prostate data set: plot of ICL and BIC values



Prostate data set: R code for prediction

```
# Choose the number of classes to study in the following.
K <- 3
resK <- extractMixtCompObject(res, K)

# Prediction
Patient1=data[1,]
Patient1["Age"]="?"
Patient1["Wt"]="[70:80]"
Patient1["EKG"]="4,5"
resPredict <- mixtCompPredict(Patient1,resLearn=resK)

# See output of prediction
resPredict$variable$data$z_class$stat
resPredict$variable$data$z_class$completed
resPredict$variable$data$Age$completed
resPredict$variable$data$Age$stat
resPredict$variable$data$Wt$completed
resPredict$variable$data$Wt$stat
resPredict$variable$data$EKG$completed
resPredict$variable$data$EKG$stat
```

Prostate data set: output results from Patient 1

```
> resPredict$variable$data$z_class$stat
      k: 1 k: 2 k: 3
[1,] 0.69   0 0.31
> resPredict$variable$data$z_class$completed
[1] 1
> resPredict$variable$data$Age$completed
[1] 70.70071
> resPredict$variable$data$Age$stat
index  median q 2.500000% q 97.500000%
      1 70.70071  55.20562  85.44931
> resPredict$variable$data$Wt$completed
[1] 76.78755
> resPredict$variable$data$Wt$stat
index  median q 2.500000% q 97.500000%
      1 76.78755  71.3697  79.84079
> resPredict$variable$data$EKG$completed
[1] 5
> resPredict$variable$data$EKG$stat
$'1'
  modality probability
      5           0.81
      4           0.19
```

Compare to true values of Patient 1...

Outline

- 1 Mixture models as a multi-purpose tool
- 2 Gaussian case
- 3 Mixed data case
- 4 Ranking data case
- 5 RMixtComp in practice
- 6 Rankcluster in practice**

Rankcluster: pitch

- Implementation of the ISR model-based clustering algorithm for ranking data
- Multivariate rankings as well as partial rankings are taken into account
- This algorithm is based on an extension of the Insertion Sorting Rank (ISR) model for ranking data, which is a meaningful and effective model parametrized by a position parameter (the modal ranking) and a dispersion parameter
- The heterogeneity of the rank population is modelled by a mixture of ISR, whereas conditional independence assumption is considered for multivariate rankings
- Available on the CRAN repository:
<https://cran.r-project.org/web/packages/Rankcluster/index.html>

Rankcluster: format of data and big4 data set

Multivariate Ranks

For multivariate ranks, the different variables are combined by column and an extra parameter (n) indicates the size of each dimension.

```
data(big4)
head(big4$data)
#>      A.uefa E.uefa C.uefa D.uefa A.pl B.pl C.pl D.pl
#> 1992-1993  1     2     3     4     1     2     3     4
#> 1993-1994  1     3     2     4     1     3     2     4
#> 1994-1995  1     3     2     4     1     2     4     3
#> 1995-1996  1     3     2     4     1     2     3     4
#> 1996-1997  1     2     3     4     1     3     2     4
#> 1997-1998  1     3     2     4     2     3     1     4
big4$m
#> [1] 4 4
```

The `big4` dataset is composed of the rankings (in ranking notation) of the “Big Four” English football teams (A: Manchester, B: Liverpool, C: Arsenal, D: Chelsea) to the English Championship (Premier League) and according to the UEFA coefficients (statistics used in Europe for ranking and seeding teams in international competitions), from 1993 to 2013.

Each variable corresponds to the ranking of four elements, so $n = c(4, 4)$. In the data matrix, the first four columns correspond to the rankings in Premier League and the four next to the ranking according to the uefa coefficient.

Partial Missing Ranks

Rankcluster manages partial missing ranks. Missing positions are denoted by \emptyset .

For example `5 0 1 2 0` indicates that the position of the second and fifth objects are unknown.

Ranks with tied positions

Rankcluster manages tied positions in ranks. Tied positions are replaced by the lowest position they share.

For example, assume there are five objects to rank. If the output rank in ranking notation is `4 3 4 1 1`, the 1 for both the objects number 4 and 5 indicates that either object 4 is in first position and object 5 in second or object 5 in second position and object 4 in first. Then the object number 2 is in third position, then objects 1 and 3 are in fourth and fifth or fifth and fourth.

Big4 data set: ready for the practical activity with Rankcluster!

- Check that in 2001 Arsenal and Chelsea had the same UEFA coefficient and then are tied for the first ranking dimension
- Consider that in 1992-1993, we only know the rank of the winner for UEFA
- Run a clustering for $K \in \{1, \dots, 3\}$
- Check the retained K value with BIC and ICL
- Check the parameters of the mixture for $K = 2$ and the related partition

Big4 set: R code

```
install.packages("Rankcluster")
library(Rankcluster)

# see the dataset
data(big4)$data

# consider that in 1992-1993, we only know the rank of the winner for UEFA
big4$data[1,2:4]=0

# clustering
res=rankclust(big4$data,m=big4$m,K=1:3,Qsem=1000,Bsem=100,Ql=500,Bl=50,maxTry=20,run=5)

# see results: K=2 is retained
res
```

Big4 set: output for $K = 2$

```

ll= -104.6243
bic = 236.6493
icl = 238.3433
proportion: 0.370478 0.629522
mu:
      dim1      dim2
c11   1 3 2 4   1 3 2 4
c12   1 3 4 2   1 4 3 2

pi:
      dim1      dim2
c11 1.0000000 0.7707837
c12 0.6968134 0.7769508

partition:
[1] 1 1 1 1 2 1 2 2 2 2 2 1 1 2 2 2 2 2 1 2 2 2

tik:
      [,1]      [,2]
[1,] 0.9936257 0.006374303
[2,] 0.9932910 0.006709006
[3,] 0.9634028 0.036597242
[4,] 0.9931952 0.006804807
[5,] 0.0000000 1.000000000
[6,] 0.9936359 0.006364105
[7,] 0.0000000 1.000000000
[8,] 0.0000000 1.000000000
[9,] 0.0000000 1.000000000
[10,] 0.0000000 1.000000000
...

```

End of Part II