

# MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data

<https://massiccc.lille.inria.fr/>

C. Biernacki

with B. Auder, G. Celeux, J. Demont, F. Langrognet, V. Kubicki, C. Poli, J. Renault, S. Iovleff

21-22 June 2018, Workshop MixStatSeq, Paris

“Mixture models: Theory and applications”



Université  
de Lille



# Outline

- 1** Introduction
- 2 Model-based clustering
- 3 Mixmod in MASSICCC
- 4 MixtComp in MASSICCC
- 5 BlockCluster in MASSICCC
- 6 Conclusion

# MASSICCC?

massiccc.lille.inria.fr

## Massive Clustering with Cloud Computing

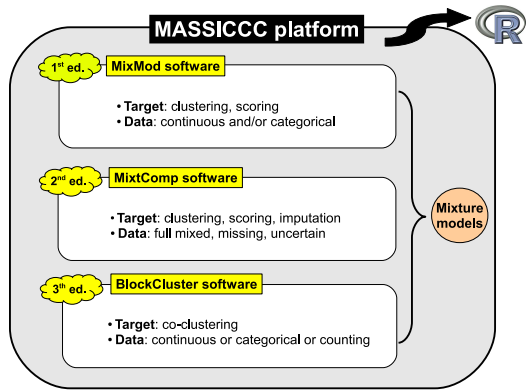
Clustering of heterogeneous data with missing values.  
Hosted in the cloud. No installation or configuration required.  
Upload your data, and get results straight away.

Developed by *Inria*

TRY IT !

SaaS: Software as a Service

# MASSICCC??



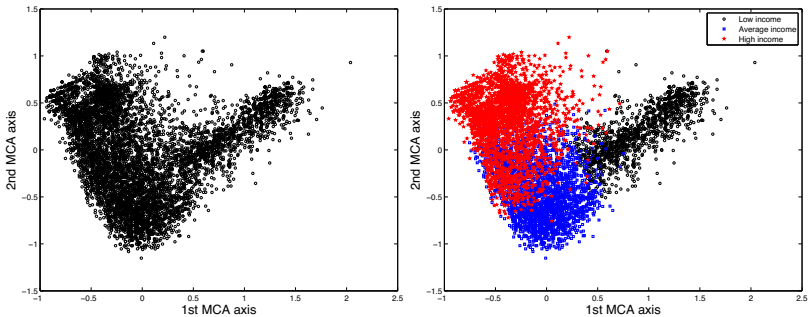
A high quality and easy to use web platform where are transferred mature research clustering (and more) software towards (non academic) professionals

Here is the computer you need!

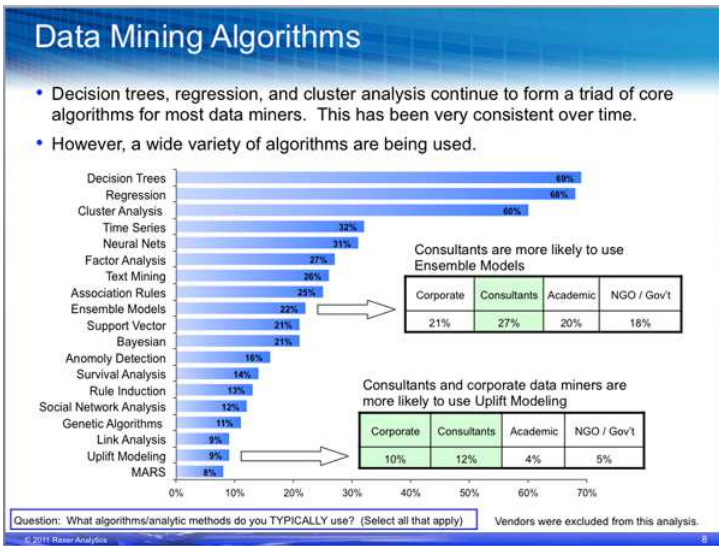


# Clustering?

Detect hidden structures in data sets

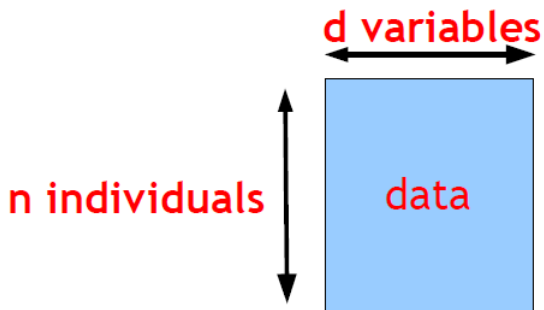


# Clustering everywhere<sup>1</sup>



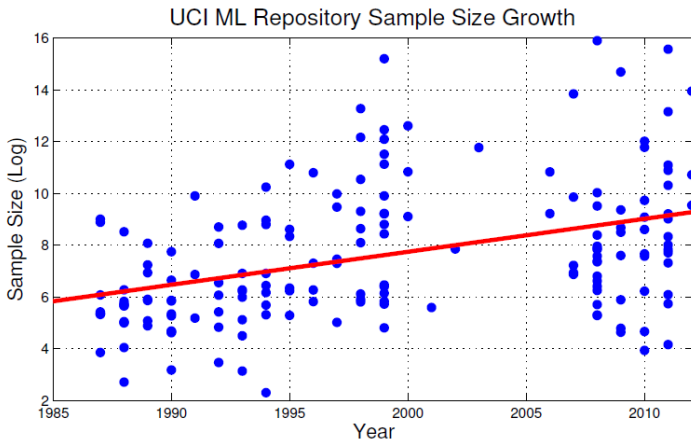
<sup>1</sup>Rexer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

## Data sets structure



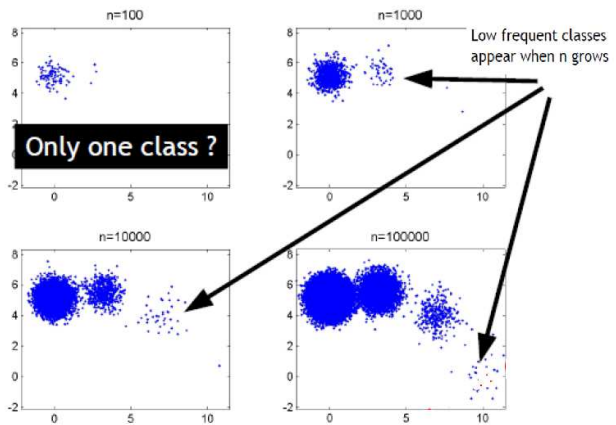


## Large data sets<sup>2</sup>



<sup>2</sup>S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, 29

# An opportunity for detecting weak signal



# Today's features: full mixed/missing



**categorical**  
Marital status  
**married**

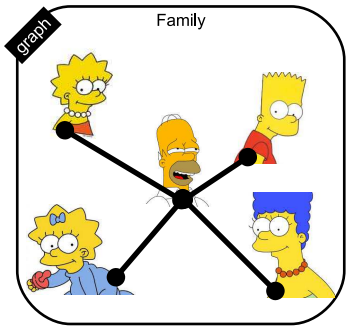
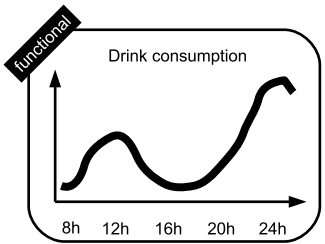
**integer**  
Children  
**3**

**missing**  
Size (m)  
**?**

**rank**  
Drink preference  
**beer > soda > water**

**ordinal**  
Intelligence  
**low**

**continuous**  
Weight (kg)  
**119.5**



**And so on...**

## Notations

- **Data:**  $n$  individuals:  $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \{\mathbf{x}^O, \mathbf{x}^M\}$  in a space  $\mathcal{X}$  of dimension  $d$ 
  - Observed individuals  $\mathbf{x}^O$
  - Missing individuals  $\mathbf{x}^M$
- **Aim:** estimation of the partition  $\mathbf{z}$  and the number of clusters  $K$   
 Partition in  $K$  clusters  $G_1, \dots, G_K$ :  $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ ,  $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$

$$\mathbf{x}_i \in G_k \Leftrightarrow z_{ih} = \mathbb{I}_{\{h=k\}}$$

### Mixed, missing, uncertain

Individuals $\mathbf{x}$				Partition $\mathbf{z}$	$\Leftrightarrow$	Group
?	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			

# Outline

- 1 Introduction
- 2 Model-based clustering**
- 3 Mixmod in MASSICCC
- 4 MixtComp in MASSICCC
- 5 BlockCluster in MASSICCC
- 6 Conclusion

## Parametric mixture model

- **Parametric assumption:**

$$p_k(\mathbf{x}_1) = p(\mathbf{x}_1; \alpha_k)$$

thus

$$p(\mathbf{x}_1) = p(\mathbf{x}_1; \theta) = \sum_{k=1}^K \pi_k p(\mathbf{x}_1; \alpha_k)$$

- **Mixture parameter:**

$$\theta = (\pi, \alpha) \text{ with } \alpha = (\alpha_1, \dots, \alpha_K)$$

- **Model:** it includes both the family  $p(\cdot; \alpha_k)$  and the number of groups  $K$

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

The number of free *continuous* parameters is given by

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

Clustering becomes a well-posed problem...

## The clustering process in mixtures

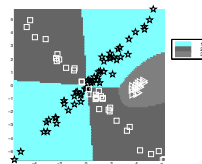
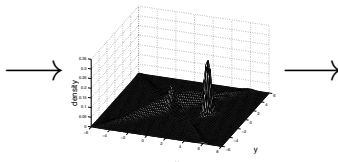
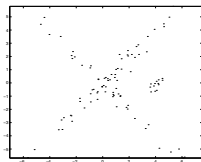
- 1 Estimation of  $\theta$  by  $\hat{\theta}$
- 2 Estimation of the **conditional probability** that  $\mathbf{x}_i \in G_k$

$$t_{ik}(\hat{\theta}) = p(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \hat{\theta}) = \frac{\hat{\pi}_k p(\mathbf{x}_i; \hat{\alpha}_k)}{p(\mathbf{x}_i; \hat{\theta})}$$

- 3 Estimation of  $z_i$  by *maximum a posteriori* (MAP)

$$\hat{z}_{ik} = \mathbb{I}_{\{k = \arg \max_{h=1, \dots, K} t_{ih}(\hat{\theta})\}}$$

- 4 **Model selection:** BIC, ICL, ...



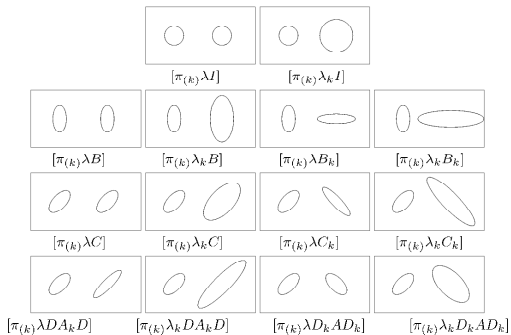
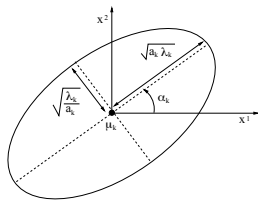
# Outline

- 1 Introduction
- 2 Model-based clustering
- 3 Mixmod in MASSICCC**
- 4 MixtComp in MASSICCC
- 5 BlockCluster in MASSICCC
- 6 Conclusion



# Only continuous features: 14 models on $\Sigma_k$

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



## Only categorical variables: latent class model

- **Categorical variables:**  $d$  variables with  $m_j$  modalities each,  $\mathbf{x}_i^j \in \{0, 1\}^{m_j}$  and

$$\mathbf{x}_i^{jh} = 1 \quad \Leftrightarrow \quad \text{variable } j \text{ of } \mathbf{x}_i \text{ takes level } h$$

- **Conditional independence:**

$$p(\mathbf{x}_i; \alpha_k) = \prod_{j=1}^d \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}$$

and

$$\alpha_k^{jh} = p(\mathbf{x}_i^{jh} = 1 | z_{ik} = 1)$$

with  $\alpha_k = (\alpha_k^{jh}; j = 1, \dots, d; h = 1, \dots, m_j)$

## Mixing continuous and categorical data: full local independence

Combine continuous and categorical data

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

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

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

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

## Estimation of $\theta$ by *complete*-likelihood

Maximize the *complete-likelihood* over  $(\theta, \mathbf{z})$

$$\ell_c(\theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \ln \{ \pi_k p(\mathbf{x}_i; \alpha_k) \}$$

- **Equivalent** to traditional methods

Metric	$\mathbf{M} = \mathbf{I}$	$\mathbf{M}$ free	$\mathbf{M}_k$ free
Gaussian model	$[\pi \lambda I]$	$[\pi \lambda C]$	$[\pi \lambda_k C_k]$

- **Bias** of  $\hat{\theta}$ : heavy if poor separated clusters
- Associated optimization algorithm: **CEM** (see later)
- CEM with  $[\pi \lambda I]$  is **strictly** equivalent to  $K$ -means
- CEM is simple et fast (convergence with few iterations)

## Estimation of $\theta$ by *observe*-likelihood

Maximize the *observe*-likelihood on  $\theta$

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^n \ln p(\mathbf{x}_i; \theta)$$

- **Convergence** of  $\hat{\theta}$ , asymptotic **efficiency**, asymptotically **unbiased**
- **General** algorithm for missing data: **EM**
- EM is simple but slower than CEM
- Interpretation: it is a kind of **fuzzy clustering**

# Principle of EM and CEM

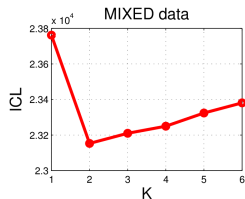
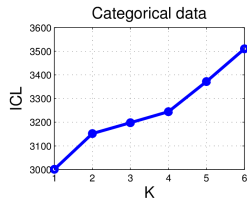
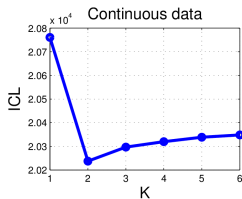
- Initialization:  $\theta^0$
- Iteration  $n^{\circ}q$ :
  - Step E: estimate probabilities  $\mathbf{t}^q = \{t_{ik}(\theta^q)\}$
  - Step C: classify by setting  $\mathbf{t}^q = \text{MAP}(\{t_{ik}(\theta^q)\})$
  - Step M: maximize  $\theta^{q+1} = \arg \max_{\theta} \ell_c(\theta; \mathbf{x}, \mathbf{t}^q)$
- Stopping rule: iteration number or criterion stability

## Properties

- $\oplus$ : simplicity, monotony, low memory requirement
- $\ominus$ : local maxima (depends on  $\theta^0$ ), linear convergence (EM)

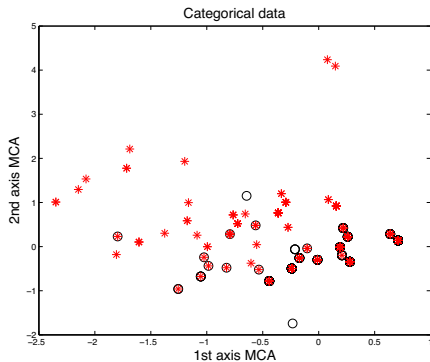
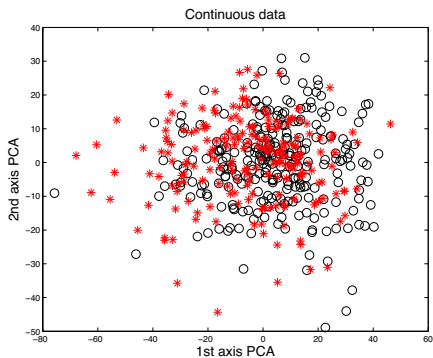
## Prostate cancer data (without mixing data)

- **Individuals:**  $n = 475$  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 histologic grade, serum prostatic acid phosphatase) and **four categorical** variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- **Model:** cond. indep.  $p(\mathbf{x}_1; \alpha_k) = p(\mathbf{x}_1; \alpha_k^{cont}) \cdot p(\mathbf{x}_1; \alpha_k^{cat})$



## Prostate cancer data (without missing data)

Variables	Continuous		Categorical		Mixed	
Error (%)	9.46		47.16		8.63	
True \ estimated group	1	2	1	2	1	2
Stage 3	247	26	142	131	252	21
Stage 4	19	183	120	82	20	182





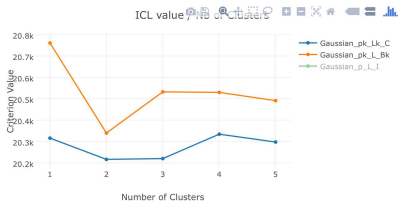
# Continuous data

## Model Criterion

This chart represents the criterion value for each model that was built. The lower the value (close to 0) the better the model.

### Criterion

ICL

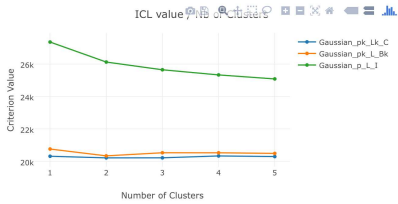


## Model Criterion

This chart represents the criterion value for each model that was built. The lower the value (close to 0) the better the model.

### Criterion

ICL



# Continuous data

MASSICCC Dashboard Help Profile Logout

RESULTS

DATA FILES

CREATE JOB

Data File: **MixMod-Example\_RFdataN.csv**

Function: **Cluster**

Labels Column:

Cluster Groups: **1-5**

**Update**

Advanced

**Outputs**

Export R Code Download Results

Samples: 475  
Variables: 8

Models

Model	Criterion	Nb Clusters	Error
Gaussian_pH_U_C	ICL(2021A/6)	2	No error
Gaussian_pH_U_C	ICL(2020/3)	2	No error
Gaussian_pH_U_C	ICL(2019/7)	5	No error
Gaussian_pH_U_C	ICL(2018A/6)	1	No error
Gaussian_pH_U_C	ICL(2034A/6)	4	No error
Gaussian_pH_L_BR	ICL(2039/8)	2	No error
Gaussian_pH_L_BR	ICL(2049L/3)	5	No error

Variables Criterion

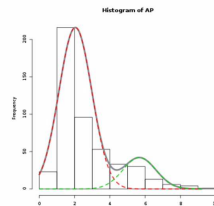
**Variable Importance**

This chart represents the discriminating level of each variable. A high value (close to one) means that the variable is highly discriminating. A low value (close to zero) means that the variable is poorly discriminating.

## Variable Parameters

This chart summarizes the distribution of the selected variable.

AP

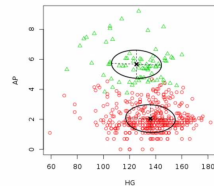


AP (Gaussian)

Show model parameters

## Biplot

HG



# Mixed data

MASSICCC Dashboard Help Profile Logout

RESULTS

DATA FILES

CREATE JOB

Title: Essai mixmod

Data File: MixMod-Example.csv

Function: Cluster

Labels Column:

Cluster Groups: 1-5

Update

Advanced

Outputs

Export R Code Download Results

Samples: 475  
Variables: 12

Models

Model	Criterion	Nb Clusters	Error
Heterogeneous_pk_Djh_Lk_Bk	ICL(23198.3)	2	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23327.2)	3	No error
Heterogeneous_pk_Djh_Lk_Bk	ICL(23402.6)	4	No error
Heterogeneous_pk_Djh_Lk_Bk	ICL(23464.3)	5	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23762.2)	1	No error

Variables Criterion

Model Criterion

This chart represents the criterion value for each model that was built. The lower the value (close to 0) the better the model.

Number of Clusters	ICL value
1	23.8k
2	23.2k
3	23.3k
4	23.4k
5	23.5k

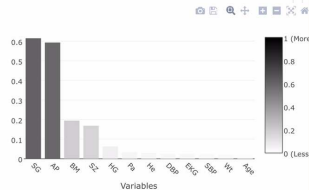
Criterion: ICL

Model	Criterion	Nb Clusters	Error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23198.3)	2	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23327.2)	3	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23402.6)	4	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23464.3)	5	No error
Heterogeneous_pk_Ekjh_Lk_Bk	ICL(23762.2)	1	No error

Variables Criterion

## Variable Importance

This chart represents the discriminating level of each variable. A high value (close to one) means that the variable is highly discriminating. A low value (close to zero) means that the variable is poorly discriminating.



## Sort Variables: II

## Variable Parameters

This chart summarizes the distribution of the selected variable.

AP

AP (Gaussian)

Hide model parameters

Class 1

mean: 3.9268, sigma: 2.8781

Class 2

mean: 1.6476, sigma: 0.2509

# Outline

- 1 Introduction
- 2 Model-based clustering
- 3 Mixmod in MASSICCC
- 4 MixtComp in MASSICCC**
- 5 BlockCluster in MASSICCC
- 6 Conclusion

## Full mixed data: conditional independence everywhere<sup>3</sup>

The aim is to combine continuous, categorical, integer data, ordinal, ranking and functional data

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

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

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

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
- ...

---

<sup>3</sup>MixtComp software on the MASSICCC platform: <https://massiccc.lille.inria.fr/>

## Missing data: MAR assumption and estimation

### Assumption on the missingness mechanism

**Missing At Random (MAR):** the probability that a variable is missing does not depend on its own value given the observed variables.

### Observed log-likelihood...

$$\ell(\boldsymbol{\theta}; \mathbf{x}^O) = \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k p(\mathbf{x}_i^O; \boldsymbol{\alpha}_k) \right) = \ln \left[ \sum_{k=1}^K \pi_k \underbrace{\int_{\mathbf{x}_i^M} p(\mathbf{x}_i^O, \mathbf{x}_i^M; \boldsymbol{\alpha}_k) d\mathbf{x}_i^M}_{\text{MAR assumption}} \right]$$

## SEM algorithm<sup>4</sup>

A SEM algorithm to estimate  $\theta$  by maximizing the **observed**-data log-likelihood

- Initialisation:  $\theta^{(0)}$
- Iteration nb  $q$ :
  - **E-step**: compute conditional probabilities  $p(\mathbf{x}^M, \mathbf{z} | \mathbf{x}^O; \theta^{(q)})$
  - **S-step**: draw  $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$  from  $p(\mathbf{x}^M, \mathbf{z} | \mathbf{x}^O; \theta^{(q)})$
  - **M-step**: maximize  $\theta^{(q+1)} = \arg \max_{\theta} \ln p(\mathbf{x}^O, \mathbf{x}^{M(q)}, \mathbf{z}^{(q)}; \theta)$
- Stopping rule: iteration number

### Properties: simpler than EM and interesting properties!

- Avoid possibly difficult E-step in an EM
- Classical M steps
- Avoids local maxima
- The mean of the sequence  $(\theta^{(q)})$  approximates  $\hat{\theta}$
- The variance of the sequence  $(\theta^{(q)})$  gives confidence intervals

---

<sup>4</sup>MixtComp software on the MASSICCC platform: <https://massiccc.lille.inria.fr/>

## Prostate cancer data (with missing data)<sup>5</sup>

- **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 histologic 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 ( $\approx 1\%$ )

We forget the classes (Stages of the disease) for performing **clustering**

### Questions

- How many clusters?
- Which partition?

---

<sup>5</sup>Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488



# Data upload without preprocessing

MASSICCC Dashboard Help Profile Logout

OVERVIEW

FILES

INPUTS

RESULTS

Age Wt PF HX SBP DBP EKG HG SZ SG A

Contli Contli Categ Categ Contli Contli Categ Contli Contli Contli

Save

Preview

	Age	Wt	PF	HX	SBP	DBP	EKG	HG	SZ	SG	AP	BM
0	75	76	1	1	15	9	5	138	1.4142	8	1.0986	1
1	76	?	?	?	?	?	?	?	5.3852	9	2.4849	?
2	54	116	1	1	13	7	4	146	6.4807	?	1.9459	1
3	69	102	1	2	14	8	5	134	1.7321	9	1.0986	1
4	66	?	?	?	?	?	?	?	1.0000	9	2.3979	?

# Run clustering analysis

MASSICCC [Dashboard](#) [Help](#) [Profile](#) [Logout](#)

OVERVIEW

FILES

INPUTS

RESULTS

## INPUTS

Parameters

Title

Data File

Package  MixMod  MixtComp  BlockCluster

Function

Labels Column  ⓘ

Cluster Groups  ⓘ

# It is running on the (Inria) cloud. . .

The screenshot shows the MASSICCC dashboard interface. At the top, there is a navigation bar with 'MASSICCC', 'Dashboard', and 'Help' on the left, and 'Profile' and 'Logout' on the right. A dark sidebar on the left contains menu items: 'OVERVIEW', 'FILES', 'INPUTS', and 'RESULTS' (which is highlighted). The main content area is titled 'RESULTS' and contains a section 'Select a job execution from the list below'. This section displays a table of three job executions:

Job ID	Job Name	Progress	Time	Status
03	Run Demo On Cancer Data Set MixComp-Example.csv	44%	5 Feb 16:59	⚙️
02	MixComp Cluster Functional-Example.csv		3 Feb 19:15	✅
01	Essai Prostate Vendredi Soir MixComp-Example.csv		3 Feb 19:03	✅

# Several quick result overviews. . . without post-processing

MASSICCC Dashboard Help Profile Logout

## Outputs

Variables 12 [Download Results](#)

Model	Criterion	Nb Clusters	Error
Default	ICL(-12239.6) BIC(-12215.2)	2	No error
Default	ICL(-12260.4) BIC(-12195.8)	3	No error
Default	ICL(-12268.6) BIC(-12208.2)	4	No error
Default	ICL(-12305.1) BIC(-12251.4)	5	No error
Default	ICL(-12375.0) BIC(-12288.9)	6	No error
Default	ICL(-12442.7) BIC(-12354.1)	7	No error
Default	ICL(-12546.1) BIC(-12546.1)	1	No error

Variable Entropy Class Entropy Parameters **Criterion Plot** Variable Similarities Class Similarities

ICL value / Nb of Clusters

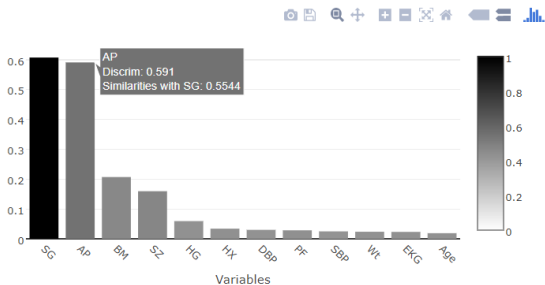
Number of Clusters	Criterion Value
1	-12.55k
2	-12.25k
3	-12.31k
4	-12.32k
5	-12.37k
6	-12.41k
7	-12.45k

# Variable significance on global partition

## Variable Importance

This chart represents the **discriminating** level of each variable. A high value (close to one) means that the variable is highly discriminating. A low value (close to zero) means that the variable is poorly discriminating. Click on one of the bars to display the distribution of this variable and, to also display the similarities between this variable and all the others. The color of the bars reflects the similarities between all the variables and the selected variable.

[Read more](#)



Sort Variables:

+ similarity between variables

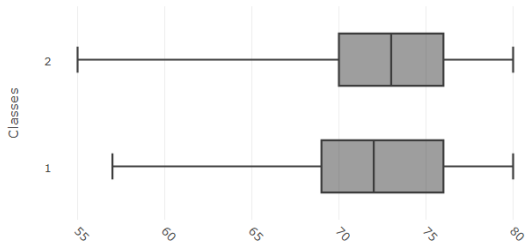
# Variable “Age” difference between clusters

## Variable Parameters

This chart summarizes the distribution of the selected variable.

Age

Boxplot of the distribution per class for Age



Age (Gaussian)

▼ Hide model parameters

Class 1

Class 2

mean: 71.534, sigma: 6.760    mean: 71.313, sigma: 7.463

# Variable “SG” difference between clusters

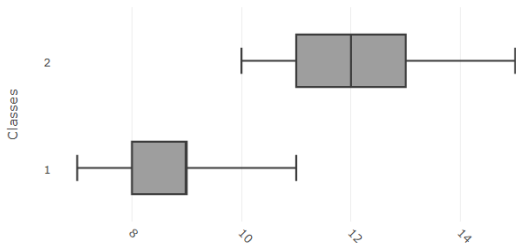


## Variable Parameters

This chart summarizes the distribution of the selected variable.

SG

Boxplot of the distribution per class for SG



SG (Gaussian)

▼ Hide model parameters

Class 1

mean: 8.940, sigma: 1.154

Class 2

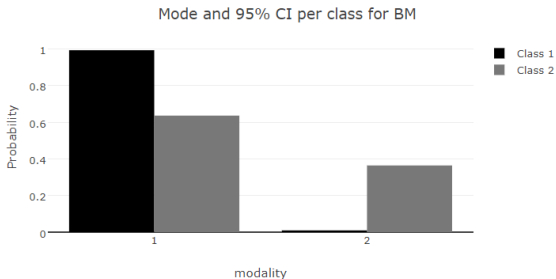
mean: 12.087, sigma: 1.405

# Variable “BM” difference between clusters

## Variable Parameters

This chart summarizes the distribution of the selected variable.

BM



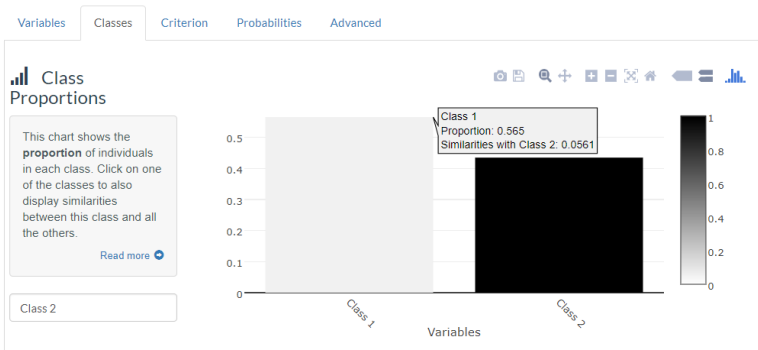
BM (Multinomial)

▼ Hide model parameters

Class 1	Class 2
scatter: [0.993,0.007]	scatter: [0.633,0.367]



# Individual cluster separation (with the cluster weight)



## Two strategies in competition

- **Strategy “mice<sup>6</sup> + MixtComp”**: MixtComp on the dataset completed by mice

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

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

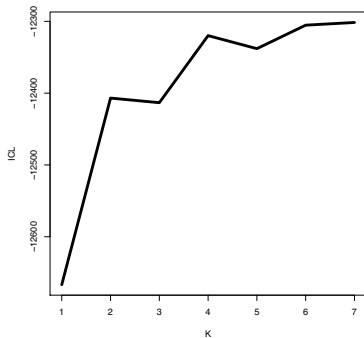
Partition quality with  $K = 2$

Strategy	mice + MixtComp	full MixtComp
% misclassified	12.8	8.1

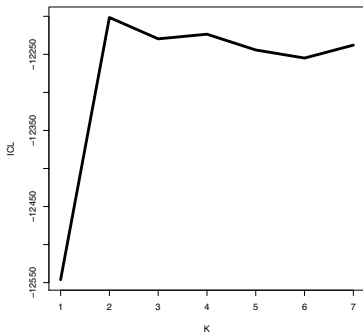
---

<sup>6</sup><http://cran.r-project.org/web/packages/mice/mice.pdf>

## Choosing $K$ with the ICL criterion



mice + MixtComp  
 $\hat{K} = 7$



full MixtComp  
 $\hat{K} = 2$

... may lose some cluster information when imputation before clustering

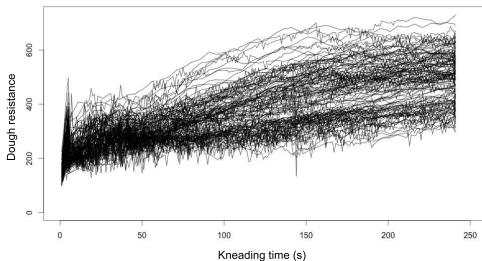
# Scoring cancer data following the clustering task

The screenshot displays the MASSICCC web interface. At the top, there is a navigation bar with 'MASSICCC', 'Dashboard', and 'Help' on the left, and 'Profile' and 'Logout' on the right. A dark sidebar on the left contains a menu with 'OVERVIEW', 'FILES', 'INPUTS' (highlighted), and 'RESULTS'. The main content area is titled 'INPUTS' and contains a 'Parameters' section. This section includes a 'Title' field with the text 'Scoring following the clustering task', a 'Data File' dropdown menu set to 'MixtComp-Example.csv', and a 'Package' selection area with radio buttons for 'MixMod', 'MixtComp' (selected), and 'BlockCluster'. Below this is a 'Function' dropdown menu with 'Predict' selected. The 'Classification Model' section shows a list of models with columns for an ID, the model name, and the creation time. A green 'Create' button is located at the bottom of the model list.

Classification Model	ID	Model Name	Creation Time
	03	Avril LEMO LN Cancer Data set MixtComp-Example.csv	5 Feb 16:59
	02	MixtComp Cluster Functional-Example.csv	3 Feb 19:15
	01	Essai Prostate Vendredi Soir MixtComp-Example.csv	3 Feb 19:03

## Curve “cookies” data set

The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process<sup>7</sup>. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval  $[0; 480]$ . The 115 flours produce cookies of different quality: 50 of them have produced cookies of good quality, 25 produced medium quality and 40 low quality.



---

<sup>7</sup>Lévêder *et al*, 04

# Upload curves data

MASSICCC [Dashboard](#) [Help](#) [Profile](#) [Logout](#)

OVERVIEW

FILES

INPUTS

RESULTS

Set:  All  ex: 0, 1, 3, 4 As: Categorical

Upload a file with a list of datatypes for each column. [?](#)

Aucun fichier choisi

Function

Preview

	Function
0	0.251.226202169594.2.257.61097125343.4.263.758...
1	0.241.129520478231.2.245.716088727869.4.250.18...
2	0.194.07006418218.2.196.01311806268.4.197.956...
3	0.137.021447956417.2.154.635389904923.4.170.65...
4	0.244.120130204111.2.245.627062897663.4.247.13...

# Run a clustering task with three clusters

MASSICCC Dashboard Help Profile Logout

OVERVIEW

FILES

**INPUTS**

RESULTS

## INPUTS

Parameters

Title

Data File

Package  MixMod  MixComp  BlockCluster

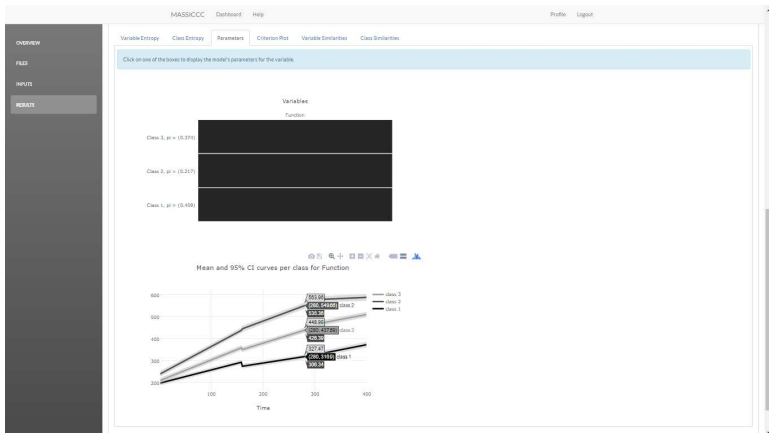
Function

Labels Column  ⓘ

Cluster Groups  ⓘ

Variable Params

# Overview of the three clusters of cookies

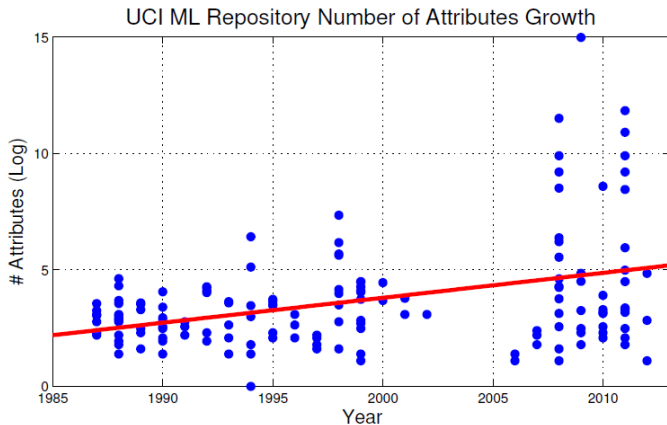




# Outline

- 1 Introduction
- 2 Model-based clustering
- 3 Mixmod in MASSICCC
- 4 MixtComp in MASSICCC
- 5 BlockCluster in MASSICCC**
- 6 Conclusion

# High-dimensional (HD) data<sup>8</sup>



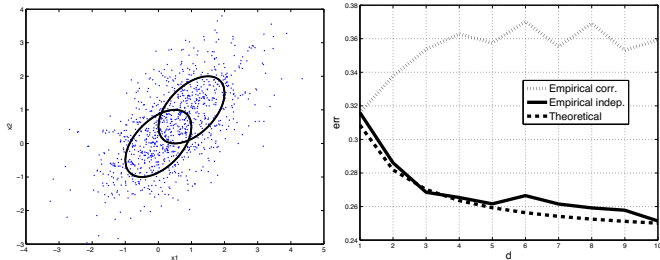
<sup>8</sup>S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, 29

## Bias/variance in HD: reduce variance, accept bias

A two-component  $d$ -variate Gaussian mixture with **intra-dependency**:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1 | z_{11} = 1 \sim N_d(\mathbf{0}, \Sigma), \quad \mathbf{X}_1 | z_{12} = 1 \sim N_d(\mathbf{1}, \Sigma)$$

- Each variable provides **equal** and **own** separation information
- Theoretical error decreases** when  $d$  grows:  $\text{err}_{\text{theo}} = \Phi(-\|\mu_2 - \mu_1\|_{\Sigma^{-1}}/2)$
- Empirical error rate with the (true) **intra-correlated model worse** with  $d$
- Empirical error rate with the (false) **intra-independent model better** with  $d$ !



## Some alternatives for reducing variance

- Dimension reduction in non-canonical space (PCA-like typically)
- Dimension reduction in the canonical space (variable selection)
- Model parsimony in the initial HD space (constraints on model parameters)

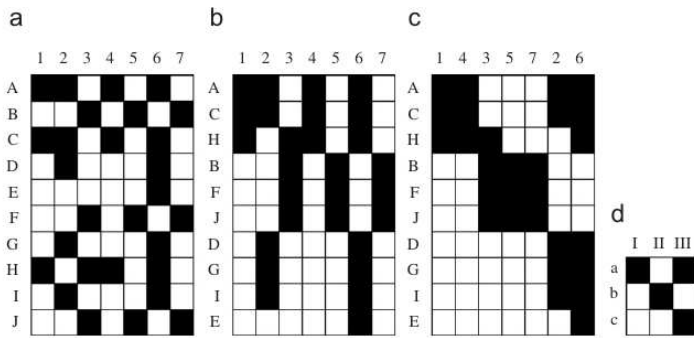
### But which kind of parsimony?

- Remember that clustering is a way for dealing with large  $n$
- Why not reusing this idea for large  $d$ ?

### Co-clustering

It performs parsimony of row clustering through variable clustering

# From clustering to co-clustering



[Govaert, 2011]

## Notations

- $\mathbf{z}_i$ : the cluster of the row  $i$
- $\mathbf{w}_j$ : the cluster of the column  $j$
- $(\mathbf{z}_i, \mathbf{w}_j)$ : the **block** of the element  $\mathbf{x}_{ij}$  (row  $i$ , column  $j$ )
  
- $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ : partition of individuals in  $K$  clusters of rows
- $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_d)$ : partition of variables in  $L$  clusters of columns
- $(\mathbf{z}, \mathbf{w})$ : **bi-partition** of the whole data set  $\mathbf{x}$
- Both space partitions are respectively denoted by  $\mathcal{Z}$  and  $\mathcal{W}$

### Restriction

All variables are of the same kind (research in progress for overcoming that. . .)

## The latent block model (LBM)

- Generalization of some existing non-probabilistic methods
- Extend the latent class principle of local (or conditional) independence
- Thus  $x_{ij}$  is assumed to be independent once  $z_i$  and  $w_j$  are fixed ( $\alpha = (\alpha_{kl})$ ):

$$p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \alpha) = \prod_{i,j} p(x_{ij}; \alpha_{z_i w_j})$$

- $\pi = (\pi_k)$  : vectors of proba.  $\pi_k$  that a row belongs to the  $k$ th row cluster
- $\rho = (\rho_l)$  : vectors of proba.  $\rho_l$  that a row belongs to the  $l$ th column cluster
- Independence between all  $z_i$  and  $w_j$
- Extension of the traditional mixture model-based clustering ( $\alpha = (\alpha_{kl})$ ):

$$p(\mathbf{x}; \theta) = \sum_{(\mathbf{z}, \mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,j} \pi_{z_i} \rho_{w_j} p(x_{ij}; \alpha_{z_i w_j})$$

## Distribution for different kinds of data

[Govaert and Nadif, 2014] The pdf  $p(\cdot; \alpha_{z_i w_j})$  depends on the kind of data  $x_{ij}$ :

- **Binary** data:  $x_{ij} \in \{0, 1\}$ ,  $p(\cdot; \alpha_{kl}) = \mathcal{B}(\alpha_{kl})$
- **Categorical** data with  $m$  levels:  
 $\mathbf{x}_{ij} = \{x_{ijh}\} \in \{0, 1\}^m$  with  $\sum_{h=1}^m x_{ijh} = 1$  and  $p(\cdot; \alpha_{kl}) = \mathcal{M}(\alpha_{kl})$  with  $\alpha_{kl} = \{\alpha_{kjh}\}$
- **Count** data:  $x_i^j \in \mathbb{N}$ ,  $p(\cdot; \alpha_{kl}) = \mathcal{P}(\mu_k \nu_l \gamma_{kl})$ <sup>9</sup>
- **Continuous** data:  $x_i^j \in \mathbb{R}$ ,  $p(\cdot; \alpha_{kl}) = \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$

---

<sup>9</sup>The Poisson parameter is here split into  $\mu_k$  and  $\nu_l$  the effects of the row  $k$  and the column  $l$  respectively and  $\gamma_{kl}$  the effect of the block  $kl$ . Unfortunately, this parameterization is not identifiable. It is therefore not possible to estimate simultaneously  $\mu_k$ ,  $\nu_l$  and  $\gamma_{kl}$  without imposing further constraints. Constraints  $\sum_k \pi_k \gamma_{kl} = \sum_l \rho_l \gamma_{kl} = 1$  and  $\sum_k \mu_k = 1, \sum_l \nu_l = 1$  are a possibility.



## Extreme parsimony ability

Model	Number of parameters
Binary	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Categorical	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL(m - 1)$
Contingency	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Continuous	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + 2KL$

Very parsimonious so well suitable for the (ultra) HD setting

$$\text{nb. param.}_{\text{HD}} = \text{nb. param.}_{\text{classic}} \times \frac{L}{d}$$

**Other advantage:** stay in the canonical space thus meaningful for the end-user

## Binary illustration: easy interpretation

[Govaert, 2011]

	<i>abcdeghij</i>
y1	1010001101
y2	0101110011
y3	1000001100
y4	1010001100
y5	0111001100
y6	0101110101
y7	0111110111
y8	1100111011
y9	0100110000
y10	1010101101
y11	1010001100
y12	1010000100
y13	1010001101
y14	0010011100
y15	0010010100
y16	1111001100
y17	0101110011
y18	1010011101
y19	1010001000
y20	1100101100

Données

Indep. B(0.83)

	<i>a c g h</i>	<i>b d e f i j</i>
y2	0 0 0 0	1 1 1 1 1 1
y6	0 0 0 1	1 1 1 1 0 1
y7	0 1 0 1	1 1 1 1 1 1
y8	1 0 1 0	1 0 1 1 1 1
y9	0 0 0 0	1 0 1 1 0 0
y17	0 0 0 0	1 1 1 1 1 1
y1	1 1 1 1	0 0 0 0 0 1
y3	1 0 1 1	0 0 0 0 0 0
y4	1 1 1 1	0 0 0 0 0 0
y5	0 1 1 1	1 1 0 0 0 0
y10	1 1 1 1	0 0 1 0 0 1
y11	1 1 1 1	0 0 0 0 0 0
y12	1 1 0 1	0 0 0 0 0 0
y13	1 1 1 1	0 0 0 0 0 1
y14	0 1 1 1	0 0 0 1 0 0
y15	0 1 0 1	0 0 0 1 0 0
y18	1 1 1 1	1 1 0 0 0 0
y18	1 1 1 1	0 0 0 1 0 1
y19	1 1 1 0	0 0 0 0 0 0
y20	1 0 1 1	1 0 1 0 0 0

Matrice réorganisée

mode

0	1
1	0

Résumé

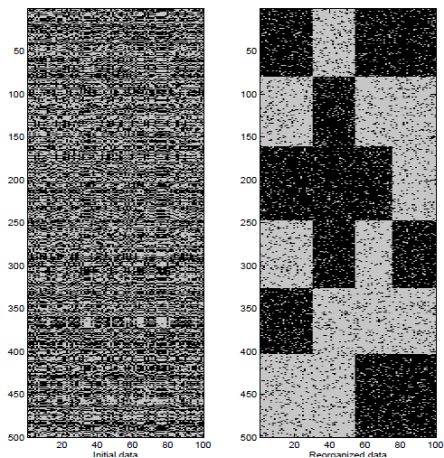
0.86	0.79
0.83	0.86

Homogénéité

proba = mode

## Binary illustration: user-friendly visualization

[Govaert, 2011]



$$n = 500, d = 10, K = 6, L = 4$$

## MLE estimation: log-likelihood(s)

- Remember Lesson 3: first estimate  $\theta$ , then deduce estimate of  $(\mathbf{z}, \mathbf{w})$
- Observed log-likelihood:  $\ell(\theta; \mathbf{x}) = \ln p(\mathbf{x}; \theta)$
- MLE:

$$\hat{\theta} = \arg \max_{\theta} \ell(\theta; \mathbf{x})$$

- Complete log-likelihood:

$$\begin{aligned} \ell_c(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) &= \ln p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \theta) \\ &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{k,l} w_{jl} \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log p(x_i^j; \alpha_{kl}) \end{aligned}$$

Be careful with asymptotics...

If  $\ln(d)/n \rightarrow 0$ ,  $\ln(n)/d \rightarrow 0$  when  $n \rightarrow \infty$  and  $d \rightarrow \infty$ , then the MLE is consistent  
 [Brault et al., 2017]

## MLE estimation: EM algorithm

- **E-step** of EM (iteration  $q$ ):

$$\begin{aligned}
 Q(\theta, \theta^{(q)}) &= E[\ell_c(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) | \mathbf{x}; \theta^{(q)}] \\
 &= \sum_{i,k} \underbrace{p(z_i = k | \mathbf{x}; \theta^{(q)})}_{t_{ik}^{(q)}} \ln \pi_k + \sum_{j,l} \underbrace{p(w_j = l | \mathbf{x}; \theta^{(q)})}_{s_{jl}^{(q)}} \ln \rho_l \\
 &\quad + \sum_{i,j,k,l} \underbrace{p(z_i = k, w_j = l | \mathbf{x}; \theta^{(q)})}_{e_{ijkl}^{(q)}} \ln p(x_{ij}; \alpha_{kl})
 \end{aligned}$$

- **M-step** of EM (iteration  $q$ ): classical. For instance, for the Bernoulli case, it gives

$$\pi_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}}{n}, \quad \rho_l^{(q+1)} = \frac{\sum_j s_{jl}^{(q)}}{d}, \quad \alpha_{kl}^{(q+1)} = \frac{\sum_{i,j} e_{ijkl}^{(q)} x_{ij}}{\sum_{i,j} e_{ijkl}^{(q)}}$$

## MLE: intractable E step

$e_{ijkl}^{(q)}$  is usually intractable. . .

- Consequence of dependency between  $x_{ij}$ 's (link between rows and columns)
- Involve  $K^n L^d$  calculus (number of possible blocks)
- Example: if  $n = d = 20$  and  $K = L = 2$  then  $10^{12}$  blocks
- Example (cont'd): 33 years with a computer calculating 100,000 blocks/second

### Alternatives to EM

- **Variational EM** (numerical approx.): conditional independence assumption

$$p(\mathbf{z}, \mathbf{w} | \mathbf{x}; \boldsymbol{\theta}) \approx p(\mathbf{z} | \mathbf{x}; \boldsymbol{\theta}) p(\mathbf{w} | \mathbf{x}; \boldsymbol{\theta})$$

- **SEM-Gibbs** (stochastic approx.): replace E-step by a S-step approx. by Gibbs

$$\mathbf{z} | \mathbf{x}, \mathbf{w}; \boldsymbol{\theta} \quad \text{and} \quad \mathbf{w} | \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}$$

## MLE: variational EM (1/2)

- Use a general variational result from [Hathaway, 1985]
- Maximizing  $\ell(\boldsymbol{\theta}; \mathbf{x})$  on  $\boldsymbol{\theta}$  is equivalent to maximize  $\tilde{\ell}_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{e})$  on  $(\boldsymbol{\theta}, \mathbf{e})$

$$\tilde{\ell}_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{e}) = \sum_{i,k} t_{ik} \ln \pi_k + \sum_{j,l} s_{jl} \ln \rho_l + \sum_{i,j,k,l} e_{ijkl} \ln p(x_{ij}; \boldsymbol{\alpha}_{kl})$$

where  $\mathbf{e} = (e_{ijkl})$ ,  $e_{ijkl} \in \{0, 1\}$ ,  $\sum_{k,l} e_{ijkl} = 1$ ,  $t_{ik} = \sum_{j,l} e_{ijkl}$ ,  $s_{jl} = \sum_{i,k} e_{ijkl}$

- Of course maximizing  $\ell(\boldsymbol{\theta}; \mathbf{x})$  or  $\tilde{\ell}_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{e})$  are both intractable
- Idea: restriction on  $\mathbf{e}$  to obtain tractability  $e_{ijkl} = t_{ik}s_{jl}$
- New variables are thus now  $\mathbf{t} = (t_{ik})$  and  $\mathbf{s} = (s_{jl})$
- As a consequence, it is a maximization of a lower bound of the max. likelihood

$$\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}) \geq \max_{\boldsymbol{\theta}, \mathbf{t}, \mathbf{s}} \tilde{\ell}_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{e})$$

## MLE: variational EM (2/2)

### Approximated E-step

$$Q(\theta, \theta^{(q)}) \approx \sum_{i,k} t_{ik}^{(q)} \ln \pi_k + \sum_{j,l} s_{jl}^{(q)} \ln \rho_l + \sum_{i,j,k,l} t_{ik}^{(q)} s_{jl}^{(q)} \ln p(x_{ij}; \alpha_{kl})$$

- We called it now VEM
- Also known as **mean field** approximation
- **Consistency** of the variational estimate [Brault *et al.*, 2017]



## MLE: local maxima

- More local maxima than in classical mixture models
- It is a consequence of many more latent variables (blocks)
- Thus: either many VEM runs, or use the SEM-Gibbs algorithm

## MLE: SEM-Gibbs

- We have already seen the SEM algorithm in Lesson 3 (thus we do not detail more)
- It limits dependency to starting point, so it limits local maxima
- The S-step: a draw  $(\mathbf{z}^{(q)}, \mathbf{w}^{(q)}) \sim p(\mathbf{z}, \mathbf{w} | \mathbf{x}; \boldsymbol{\theta}^{(q)})$  instead an expectation
- But it is still intractable, thus use a Gibbs algorithm to approx. this draw

### Approximated S-step

Two easy draws

$$\mathbf{z}^{(q)} \sim p(\mathbf{z} | \mathbf{w}^{(q-1)}, \mathbf{x}; \boldsymbol{\theta}^{(q)})$$

and

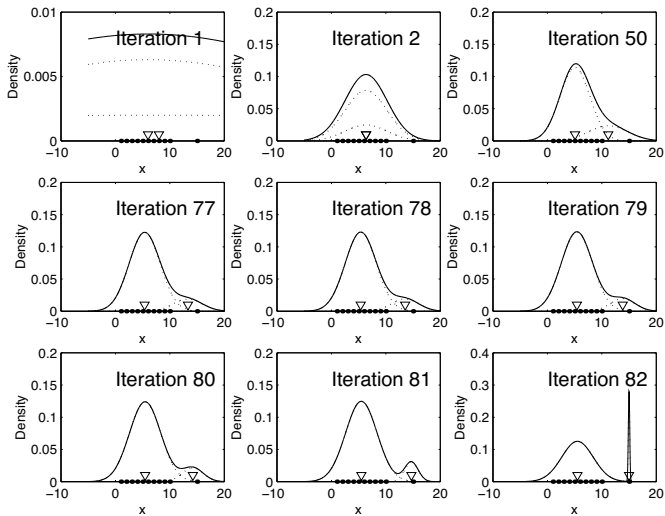
$$\mathbf{w}^{(q)} \sim p(\mathbf{w} | \mathbf{z}^{(q)}, \mathbf{x}; \boldsymbol{\theta}^{(q)})$$

- Rigorously speaking, many draws within the S-step should be performed
- Indeed, Gibbs has to reach a stochastic convergence
- In practice it works well while saving computation time

## MLE: degeneracy

- More degenerate situations than in classical mixture models
- It is again a consequence of many more latent variables (blocks)
- The Bayesian regularization (instead MLE) can be an answer

## Illustration of a degenerate situation



## Bayesian estimation: pitch

- Everything passes by the **posterior distribution of  $\theta$**

$$p(\theta|\mathbf{x}) \propto \underbrace{p(\mathbf{x}|\theta)}_{\text{log-likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

- Then, take (for instance) the **MAP** as a  $\theta$  estimate (use a VEM like algo...)

$$\hat{\theta} = \arg \max_{\theta} p(\theta|\mathbf{x})$$

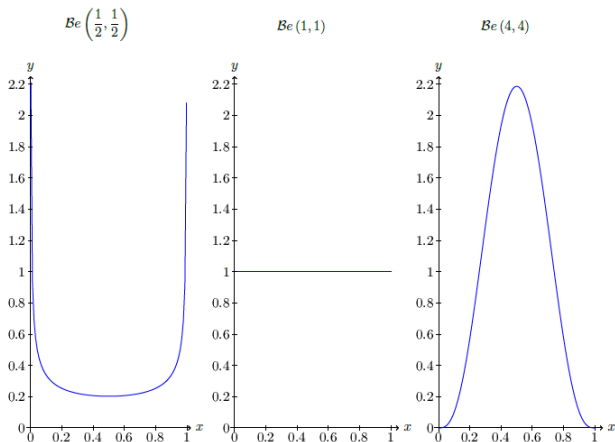
## Bayesian estimation: limiting degeneracy

- Interest for avoiding degeneracy is the prior: it acts as a **penalization** term
- Typical choices are **Dirichlet** for  $\pi$  and  $\rho$  (with independence between  $\pi$ ,  $\rho$ ,  $\alpha$ )

$$p(\theta) = \underbrace{p(\pi)}_{D_K(a, \dots, a)} \times \underbrace{p(\rho)}_{D_L(a, \dots, a)} \times \underbrace{p(\alpha)}_{\text{model dependent}}$$

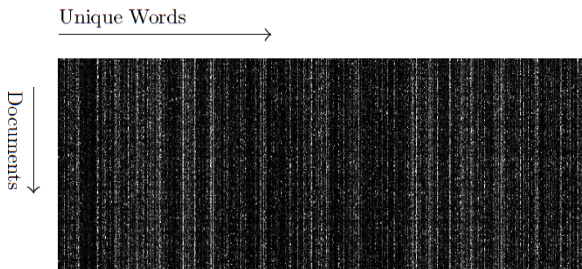
- The Dirichlet distribution is conjugate, thus easy calculus
- **Control degeneracy frequency with the  $a$  value:**
  - $a = 1$ : uniform prior, so  $\hat{\theta}$  is strictly the MLE (no regularisation)
  - $a = 1/2$ : Jeffreys prior, classical (no informative prior) but may favor degeneracy
  - $a = 4$ : a rule of thumb working well for limiting degeneracy frequency

## Bayesian estimation: prior overview



## Document clustering (1/2)

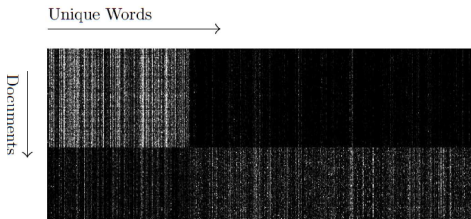
- Mixture of 1033 medical summaries and 1398 aeronautics summaries
- **Lines:** 2431 documents
- **Columns:** present words (except stop), thus 9275 unique words
- Data matrix: cross counting document  $\times$  words
- Poisson model





## Document clustering (2/2)

$$p(\hat{\mathbf{z}} \neq \mathbf{z}) \leq 2n \exp \left\{ -\frac{1}{8} d \left[ \underbrace{\min_{k \neq k'} |\tau_k - \tau_{k'}|}_{\text{overlap}} \right] \right\} + K(1 - \min_k \pi_k)^n$$



### Results with $2 \times 2$ blocs

	Medline	Cranfield
Medline	1033	0
Cranfield	0	1398

# Running BlockCluster

## Configuration

If you change the configuration of your job and save it, it will start a new process with the updated parameters. This will erase previous results.

Parameters

Title	<input type="text" value="Trial BlockCluster"/>
Data File	<input type="text" value="Blockcluster-Example.csv"/>
Data Type	<input type="text" value="Categorical"/> ⓘ
Rows Cluster Groups	<input type="text" value="1:5"/> ⓘ
Column Cluster Groups	<input type="text" value="1:5"/> ⓘ

# Running BlockCluster

MASSICCC Dashboard Help Profile Logout









RESULTS

DATA FILES

CREATE JOB

RESULTS

Select a job execution from the list below

69		Trial BlockCluster Blockcluster-Example.csv	<div style="width: 42%;"><div style="width: 42%;"></div></div>	23 May 20:47 
68		Genes K1-12 log.cpm.txt		23 May 08:12 
67		Genes log.cpm.txt		22 May 15:38 
65		Genes K1-10 log.cpm.txt		22 May 15:27 

# Running BlockCluster

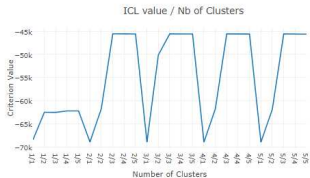
Model	Criterion	Nb Clusters	Error
<i>pik_rho_multi</i>	ICL (-45557.1)	[2,3]	No error
<i>pik_rho_multi</i>	ICL (-45563.3)	[3,3]	No error
<i>pik_rho_multi</i>	ICL (-45566.6)	[2,4]	No error
<i>pik_rho_multi</i>	ICL (-45573.9)	[4,3]	No error
<i>pik_rho_multi</i>	ICL (-45574.6)	[5,3]	No error
<i>pik_rho_multi</i>	ICL (-45577.7)	[3,4]	No error
<i>pik_rho_multi</i>	ICL (-45578.8)	[2,5]	No error

Cluster Plot

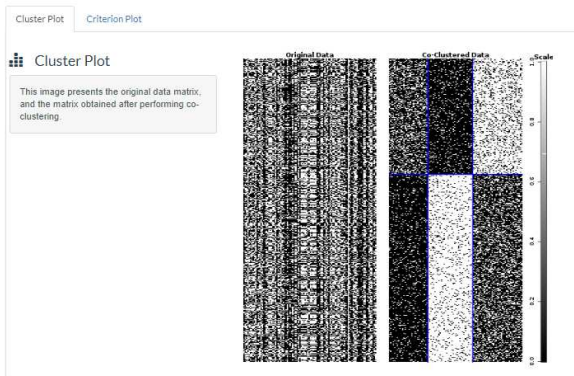
Criterion Plot

## Model Criterion

This chart represents the criterion value for each model that was built. The higher the value (close to 0) the better the model.



# Running BlockCluster



# Outline

- 1 Introduction
- 2 Model-based clustering
- 3 Mixmod in MASSICCC
- 4 MixtComp in MASSICCC
- 5 BlockCluster in MASSICCC
- 6 Conclusion**

- Use probabilistic modelling as a mathematical guideline
- Use the MASSICCC platform for user-friendly implementation

**<https://massiccc.lille.inria.fr/>**