

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"



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

The screenshot shows the homepage of the MASSICCC website. The top half features a dark background with a blurred, abstract pattern of light spots resembling a microscopic view of cells or data points. Overlaid on this are three lines of text: "Massive Clustering with Cloud Computing", "Clustering of heterogeneous data with missing values.", and "Hosted in the cloud. No installation or configuration required.". Below this, another line reads "Upload your data, and get results straight away.". The bottom half has a solid light gray background. On the left, the text "Developed by" is followed by the "Inria" logo in its signature script font. To the right of the logo is a blue rectangular button with the white text "TRY IT!".

Massive Clustering with Cloud Computing

Clustering of heterogeneous data with missing values.

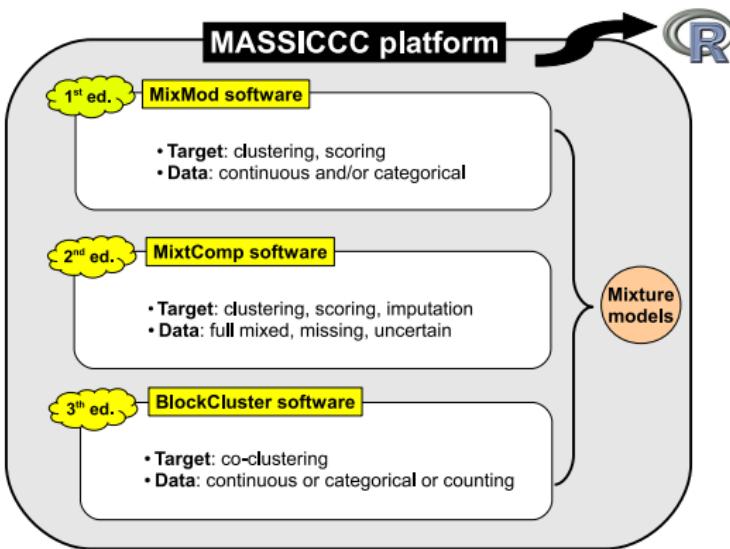
Hosted in the cloud. No installation or configuration required.

Upload your data, and get results straight away.

Developed by *Inria* TRY IT!

SaaS: Software as a Service

MASSICCC??



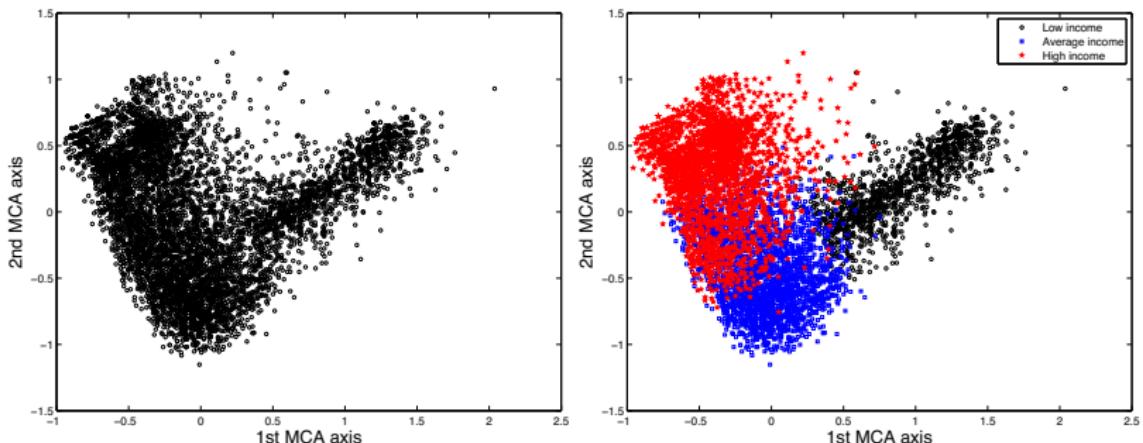
A high quality and easy to use web platform
where are transferred mature research clustering (and more) software
towards (non academic) professionals

Here is the computer you need!



Clustering?

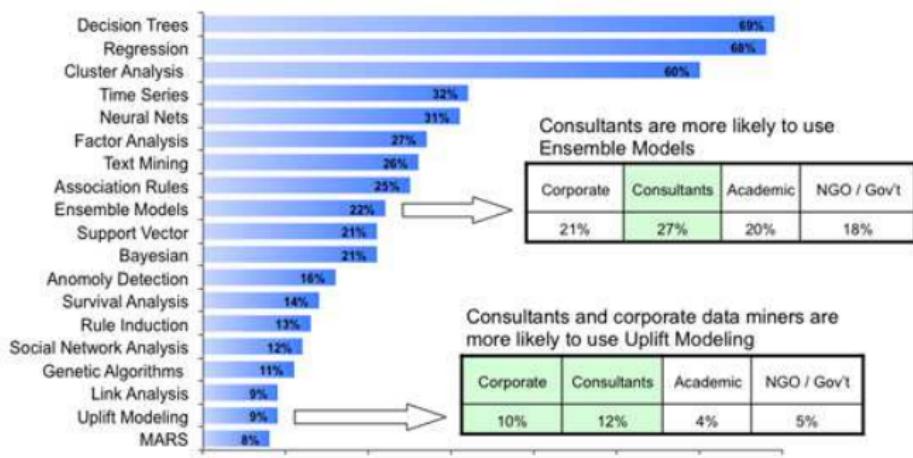
Detect hidden structures in data sets



Clustering everywhere¹

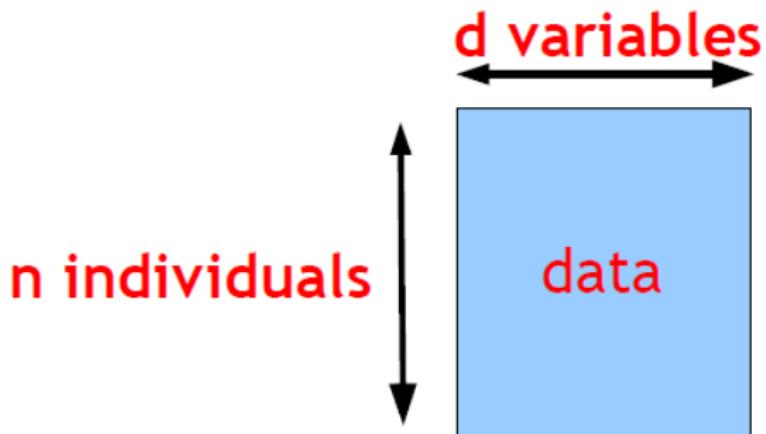
Data Mining Algorithms

- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been very consistent over time.
- However, a wide variety of algorithms are being used.

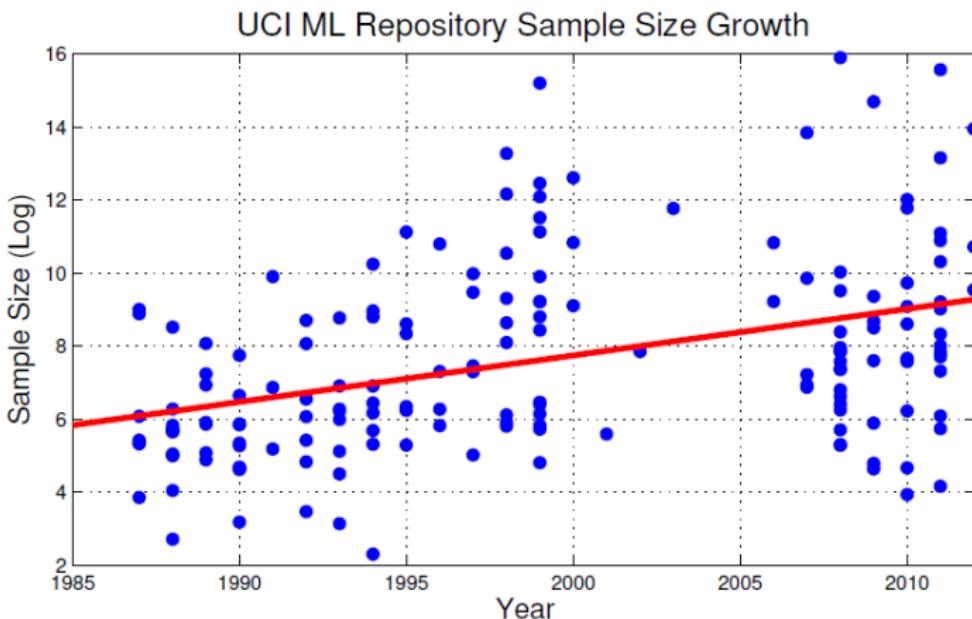


¹Rixer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

Data sets structure

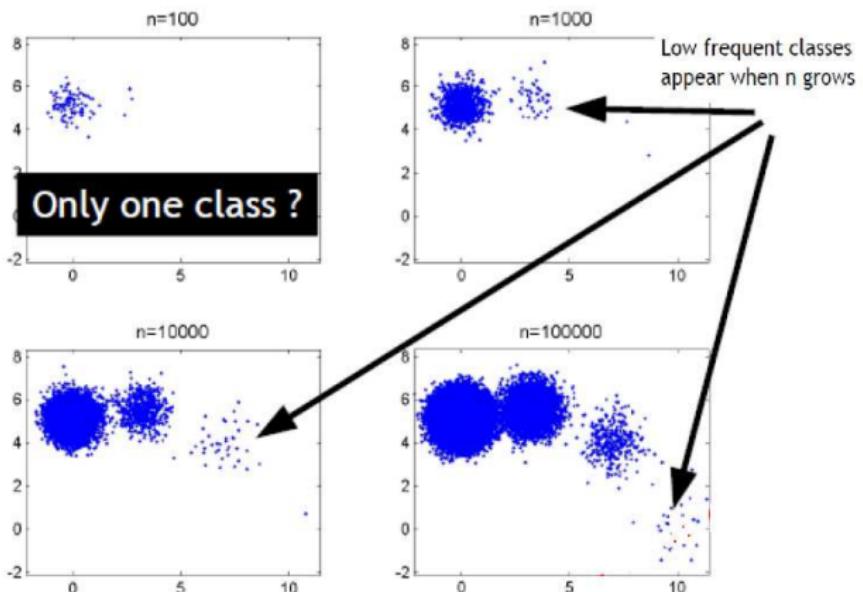


Large data sets²



²S. 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



Todays features: full mixed/missing

The slide illustrates various data types and features:

- Marital status: married** (categorical)
- Children: 3** (integer)
- Size (m): ?** (missing)
- Drink preference: beer > soda > water** (rank)
- Intelligence: low** (ordinal)
- Weight (kg): 119.5** (continuous)
- Drink consumption** (functional): A line graph showing consumption over time (8h, 12h, 16h, 20h, 24h).
- Family** (graph): A network diagram showing relationships between the Simpson family members.

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 : $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$

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

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:

$$p_k(x_1) = p(x_1; \alpha_k)$$

thus

$$p(x_1) = p(x_1; \theta) = \sum_{k=1}^K \pi_k p(x_1; \alpha_k)$$

- 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

$$\mathbf{m} = \{p(x_1; \theta) : \theta \in \Theta\}$$

The number of free *continuous* parameters is given by

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

Clustering becomes a well-posed problem...

The clustering process in mixtures

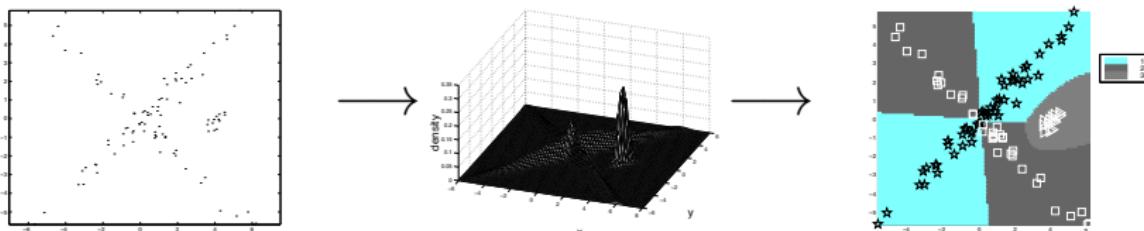
- 1 Estimation of θ by $\hat{\theta}$
- 2 Estimation of the conditional probability that $\mathbf{x}_i \in G_k$

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

- 3 Estimation of \mathbf{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

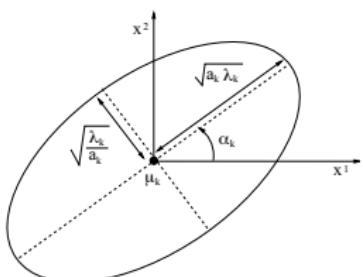
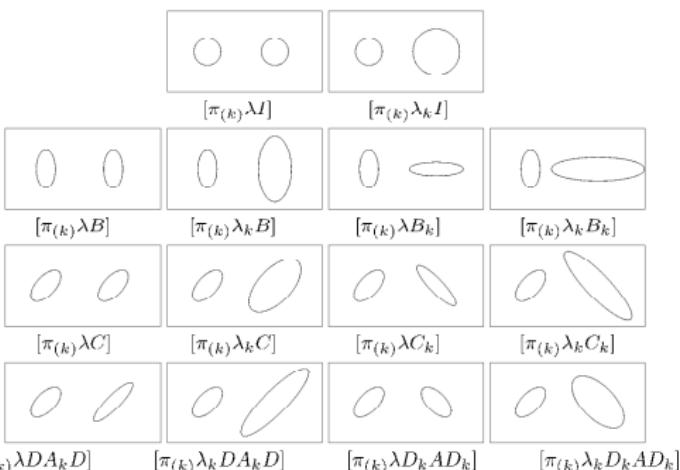
4 MixtComp in MASSICCC

5 BlockCluster in MASSICCC

6 Conclusion

Only continuous features: 14 models on Σ_k

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



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

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

- Conditional independence:

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

and

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

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

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

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

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 θ by complete-likelihood

Maximize the complete-likelihood over (θ, z)

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

- 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 θ by observe-likelihood

Maximize the observe-likelihood on θ

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

- 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

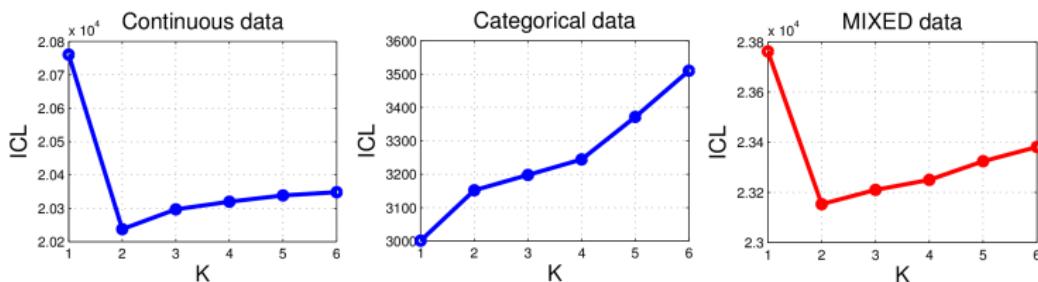
- Initialization: θ^0
- Iteration $n^o q$:
 - Step E: estimate probabilities $t^q = \{t_{ik}(\theta^q)\}$
 - Step C: classify by setting $t^q = \text{MAP}(\{t_{ik}(\theta^q)\})$
 - Step M: maximize $\theta^{q+1} = \arg \max_{\theta} \ell_c(\theta; \mathbf{x}, \mathbf{t}^q)$
- Stopping rule: iteration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

Properties

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

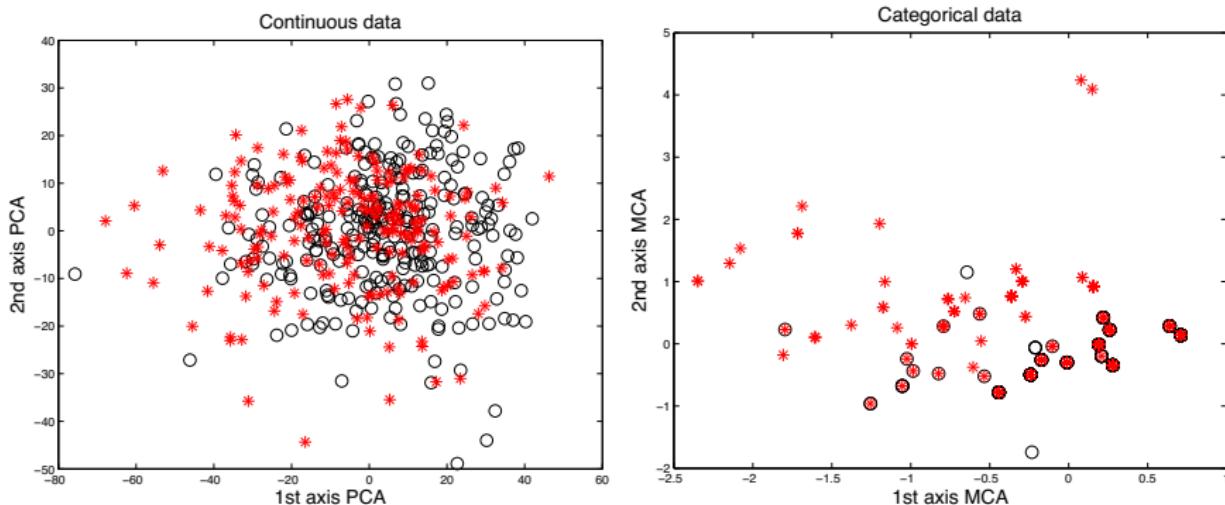
Prostate cancer data (without mixing data)

- **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. $p(\mathbf{x}_1; \boldsymbol{\alpha}_k) = p(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cont}) \cdot p(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cat})$

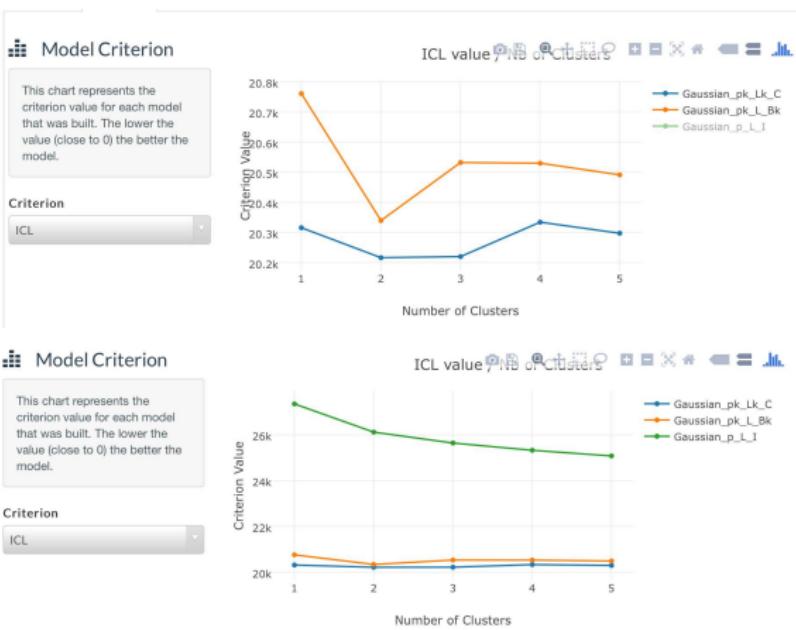


Prostate cancer data (without missing data)

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



Continuous data

MASSICCC Dashboard Help

Data File: MixMod-Example_1/FFdltN.csv

Function: Cluster

Labels Column:

Cluster Groups: 1-5

Update

Advanced

Outputs

Samples: 475 Variables: 8

Models

Model	Criterion	Nb Clusters	Error
Gaussian_pk_Ix_C	ICU(20216.0)	2	No error
Gaussian_pk_Ix_C	ICU(202020.0)	3	No error
Gaussian_pk_Ix_C	ICU(202977.7)	5	No error
Gaussian_pk_Ix_C	ICU(20016.0)	1	No error
Gaussian_pk_Ix_C	ICU(20034.0)	4	No error
Gaussian_pk_Ix_Bk	ICU(202099.0)	2	No error
Gaussian_pk_Ix_Bk	ICU(20491.0)	5	No error

Variables Criterion

Variable Importance

This chart represents the discriminating level of each variable. A high value means that one can infer 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

Histogram of AP

AP (Gaussian)

Show model parameters

Biplot

HG

AP

HG

Mixed data

MASSICCC Dashboard Help

RESULTS

Title: Essai mixmod

Data File: MixMod-Example.csv

Function: Cluster

Labels Columns:

Cluster Groups: 1-5

Update

Advanced

Outputs

Samples: 475 Variables: 12

Export R Code **Download Results**

Models

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**

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

Number of Clusters: 1, 2, 3, 4, 5

ICL value: 23.8k, 23.6k, 23.4k, 23.2k

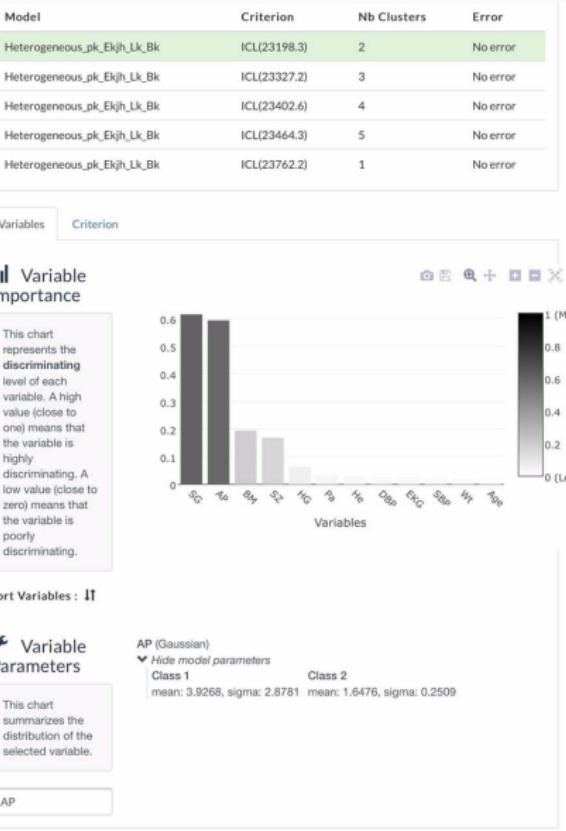
Sort Variables : IT

Variable Parameters

This chart summarizes the distribution of the selected variable.

AP (Gaussian)

AP



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 everywhere³

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**

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

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
- ...

³MixtComp 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 algorithm⁴

A SEM algorithm to estimate θ by maximizing the **observed**-data log-likelihood

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

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

Prostate cancer data (with missing data)⁵

- **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 forgot the classes (Stages of the disease) for performing clustering

Questions

- How many clusters?
- Which partition?

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

Data upload without preprocessing

The screenshot shows the MASSICCC web application interface. On the left, a dark sidebar contains navigation links: OVERVIEW, FILES (which is selected), INPUTS, and RESULTS. The main area has a header with tabs: MASSICCC, Dashboard, Help, Profile, and Logout. Below the header is a form for data entry with columns: Age, Wt, PF, HX, SBP, DBP, EKG, HG, SZ, SG, AP, and BM. Each column has a dropdown menu with options like Conti, Categ, and ? (for categorical data). A blue 'Save' button is located below the first row of columns. At the bottom, there is a 'Preview' section displaying a table of data rows:

	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

The screenshot shows the MASSICCC web application interface. On the left, there is a sidebar with tabs: OVERVIEW, FILES, INPUTS (which is currently selected), and RESULTS. The main area is titled "INPUTS". It contains a "Parameters" section with the following fields:

- Title: Run demo on cancer data set
- Data File: MixtComp-Example.csv
- Package: MixMod MixComp BlockCluster
- Function: Cluster
- Labels Column: (empty input field)
- Cluster Groups: 1-7

At the bottom of the Parameters section is a green "Create" button.

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

MASSICCC Dashboard Help Profile Logout

OVERVIEW RESULTS

FILES INPUTS

RESULTS

Select a job execution from the list below

03		Run Demo On 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	

Several quick result overviews... without post-processing

MASSICCC Dashboard Help Profile Logout Download Results

OVERVIEW FILES INPUTS RESULTS

Outputs Variables 12

Models

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	-12.21k
3	-12.20k
4	-12.22k
5	-12.30k
6	-12.38k
7	-12.45k

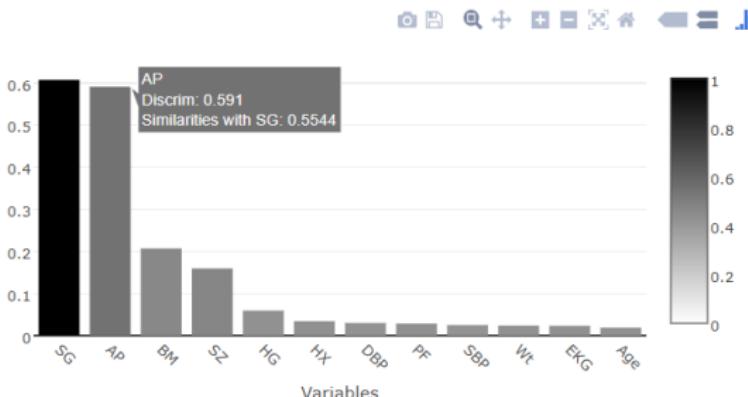
Number of Clusters

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:

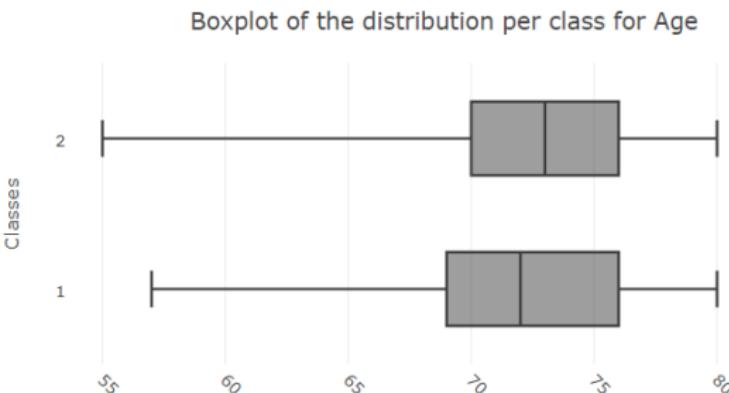
+ similarity between variables

Variable “Age” difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

Age



Age (Gaussian)

▼ Hide model parameters

Class 1

mean: 71.534, sigma: 6.760

Class 2

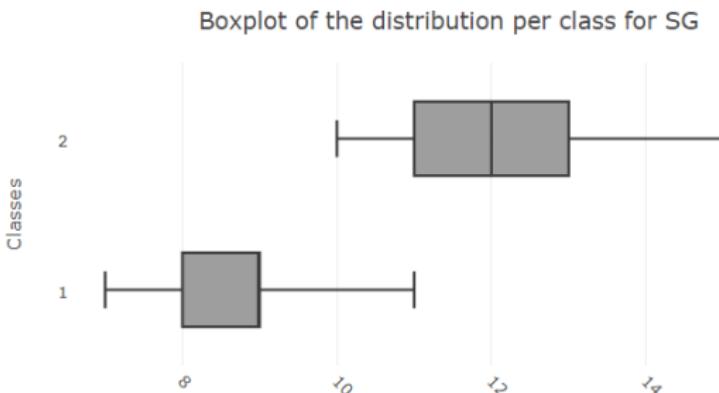
mean: 71.313, sigma: 7.463

Variable "SG" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

SG



SG (Gaussian)

▼ Hide model parameters

Class 1

mean: 8.940, sigma: 1.154

Class 2

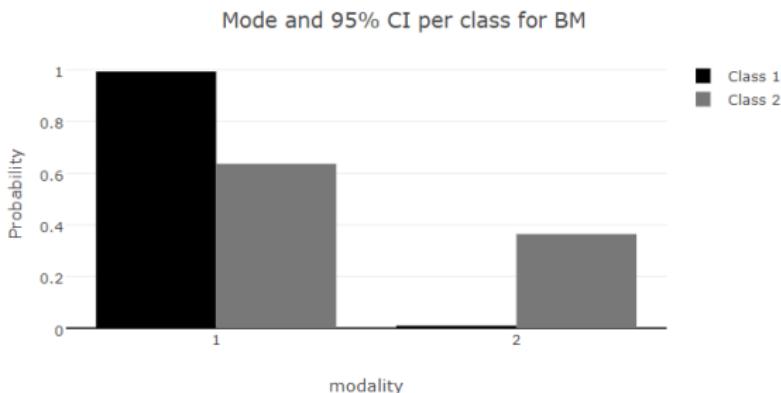
mean: 12.087, sigma: 1.405

Variable “BM” difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

BM

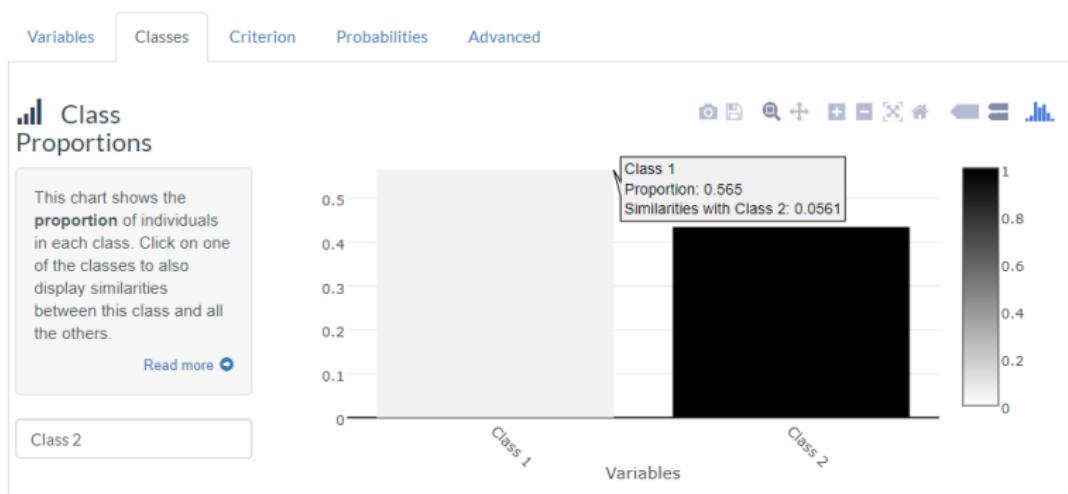


BM (Multinomial)

▼ 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 “mice⁶ + MixtComp”: MixtComp on the dataset completed by mice

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

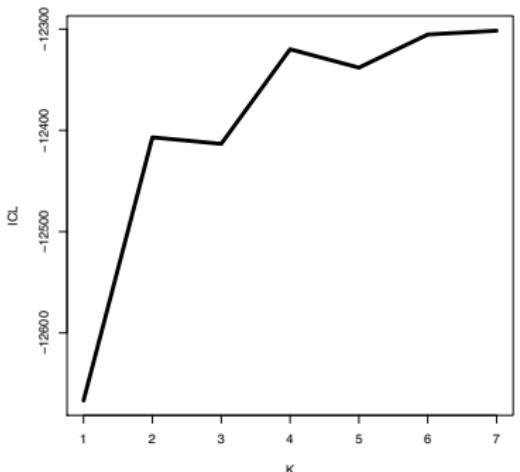
- Strategy “full MixtComp”: MixtComp on the observed (no completed) dataset

Partition quality with $K = 2$

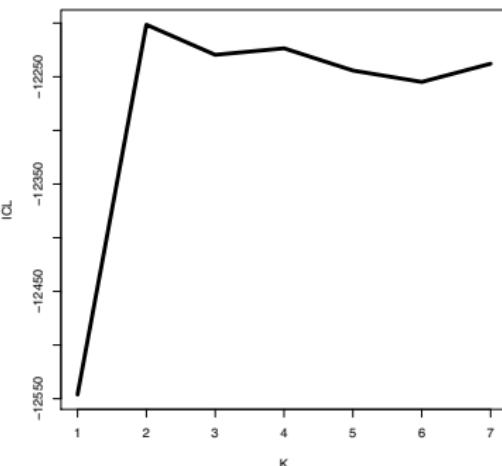
Strategy	mice + MixtComp	full MixtComp
% misclassified	12.8	8.1

⁶<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

MASSICCC Dashboard Help Profile Logout

OVERVIEW FILES INPUTS RESULTS

INPUTS

Parameters

Title: Scoring following the clustering task

Data File: MixtComp-Example.csv

Package: MixMod MixComp BlockCluster

Function: Predict

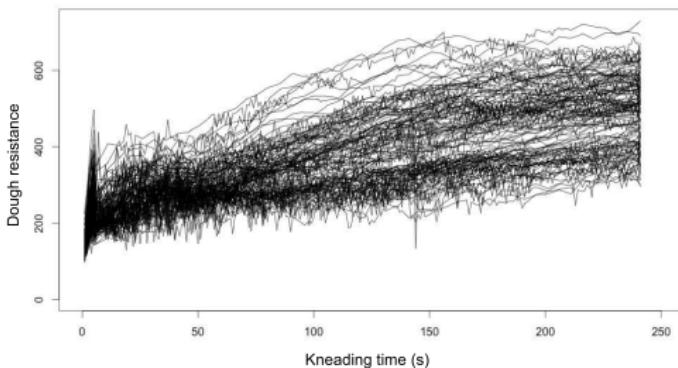
Classification Model:

Run ID	Description	Date
03	Run Demo On Cancer Data Set MixtComp-Example.csv	5 Feb 16:59
02	MixComp Cluster Functional-Example.csv	3 Feb 19:15
01	Essai Prostate Vendredi Soir MixtComp-Example.csv	3 Feb 19:03

Create

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 process⁷. 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.



⁷Lévéder et al, 04

Upload curves data

The screenshot shows the MASSICCC web application interface. On the left, there is a vertical sidebar with four tabs: 'OVERVIEW' (selected), 'FILES' (highlighted in grey), 'INPUTS', and 'RESULTS'. The main content area has a header with 'MASSICCC', 'Dashboard', and 'Help' on the left, and 'Profile' and 'Logout' on the right.

In the 'FILES' section, there is a search bar with 'Set: * All' and a placeholder 'ex: 0, 1, 3;4' with a 'As' dropdown set to 'Categorical' and an 'Apply' button. Below this is a file upload input field with the placeholder 'Choisissez un fichier' and a note 'Aucun fichier choisi'.

The 'Function' section contains a dropdown menu currently set to 'Functional' with a 'Save' button below it.

The 'Preview' section shows a table with a single column labeled 'Function' containing five rows of data:

	Function
0	0.251.226202169594.2257.61097125343.4263.758..
1	0.241.129520478231.2.245.716088727869.4250.18..
2	0.194.07006418218.2.196.013131806268.4197.956..
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: Clustering of cookies into three clusters

Data File: Functional-Example.csv

Package: MixMod (selected), MixtComp, BlockCluster

Function: Cluster

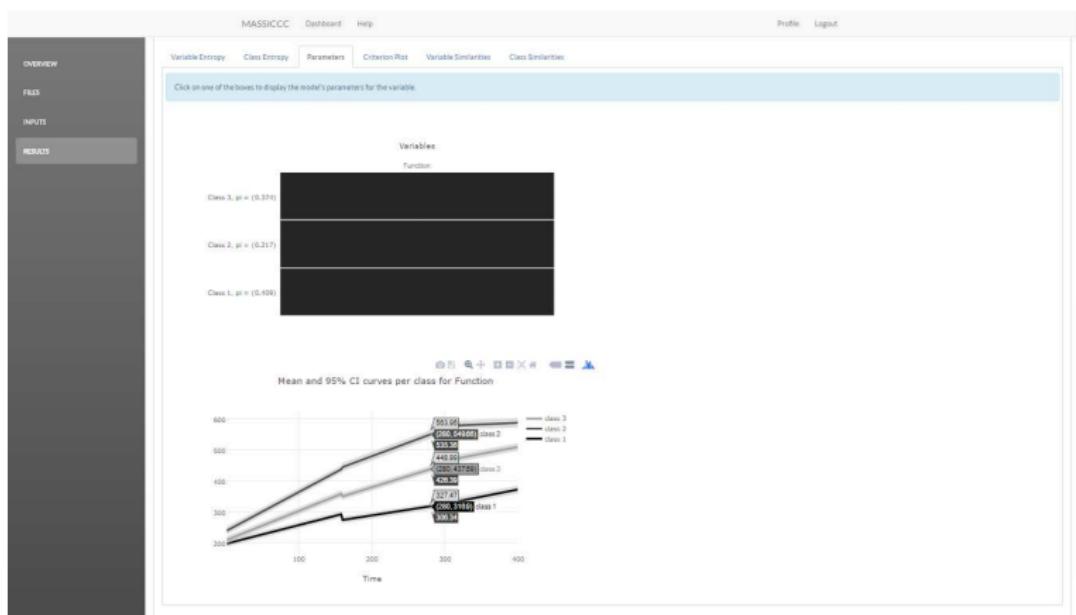
Labels Column: (empty)

Cluster Groups: 3

Variable Params: (empty)

Create

Overview of the three clusters of cookies



Outline

1 Introduction

2 Model-based clustering

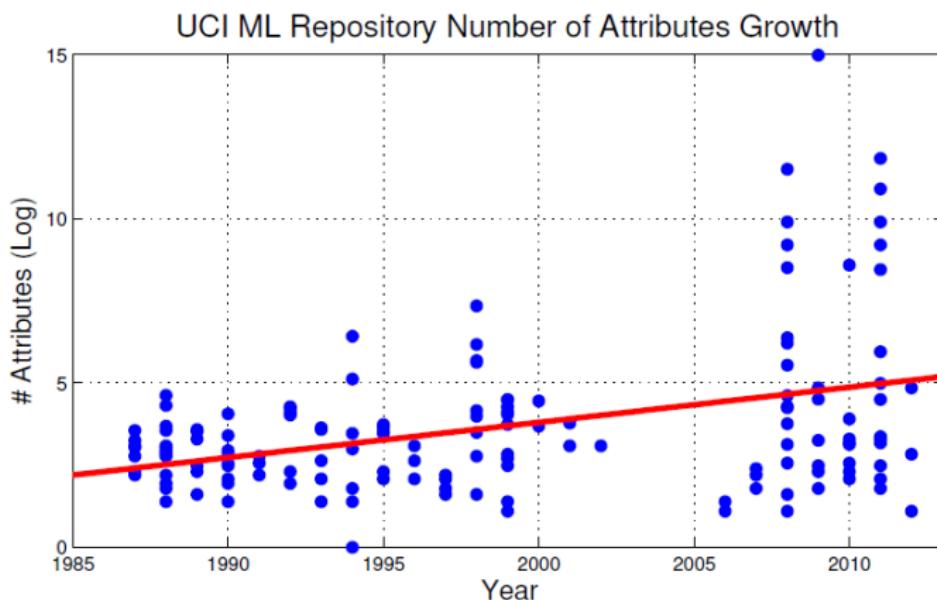
3 Mixmod in MASSICCC

4 MixtComp in MASSICCC

5 BlockCluster in MASSICCC

6 Conclusion

High-dimensional (HD) data⁸



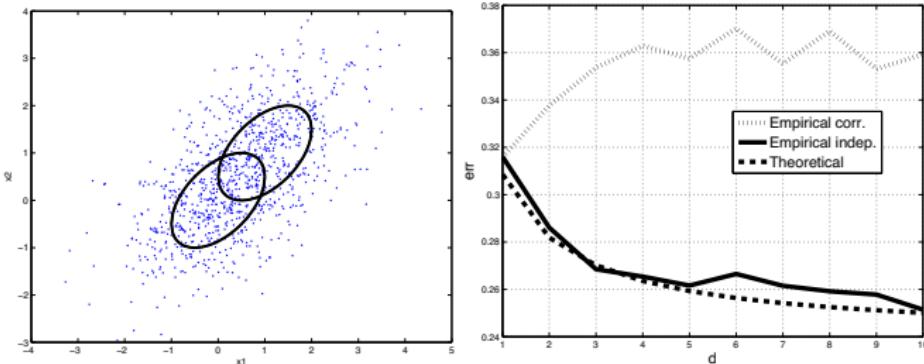
⁸S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, 29

Bias/variance in HD: reduce variance, accept bias

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

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

- 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)

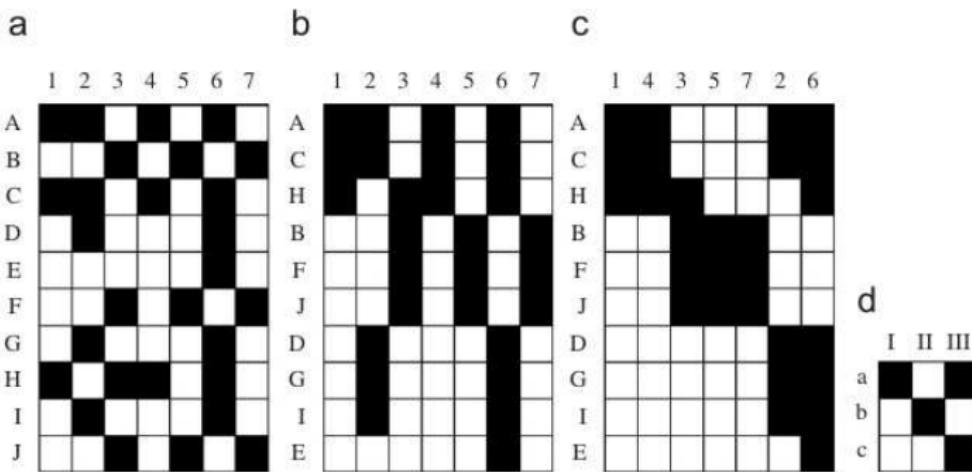
But which kind of parsimony?

- 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

- z_i : the cluster of the row i
 - w_j : the cluster of the column j
 - (z_i, w_j) : the **block** of the element x_{ij} (row i , column j)
-
- $\mathbf{z} = (z_1, \dots, z_n)$: partition of individuals in K clusters of rows
 - $\mathbf{w} = (w_1, \dots, w_d)$: 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 ($\alpha = (\alpha_{kl})$):

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

- $\pi = (\pi_k)$: vectors of proba. π_k that a row belongs to the k th row cluster
- $\rho = (\rho_l)$: vectors of proba. ρ_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 ($\alpha = (\alpha_{kl})$):

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

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

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

Extreme parsimony ability

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

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>abcdefghijklj</i>
<i>y</i> ₁	1010001101
<i>y</i> ₂	0101110011
<i>y</i> ₃	1000001100
<i>y</i> ₄	1010001100
<i>y</i> ₅	0111001100
<i>y</i> ₆	0101110101
<i>y</i> ₇	0111110111
<i>y</i> ₈	1100111011
<i>y</i> ₉	0100110000
<i>y</i> ₁₀	1010011101
<i>y</i> ₁₁	1010001100
<i>y</i> ₁₂	1010000100
<i>y</i> ₁₃	1010001101
<i>y</i> ₁₄	0010011100
<i>y</i> ₁₅	0010010100
<i>y</i> ₁₆	1111001100
<i>y</i> ₁₇	0101110011
<i>y</i> ₁₈	1010011101
<i>y</i> ₁₉	1010001000
<i>y</i> ₂₀	1100101100

Données

Indep. B(0.83)

	<i>a c g h</i>	<i>b d e f i j</i>
<i>y</i> ₂	0 0 0 0	1 1 1 1 1 1
<i>y</i> ₆	0 0 0 1	1 1 1 1 0 1
<i>y</i> ₇	0 1 0 1	1 1 1 1 1 1
<i>y</i> ₈	1 0 1 0	1 0 1 1 1 1
<i>y</i> ₉	0 0 0 0	1 0 1 1 0 0
<i>y</i> ₁₇	0 0 0 0	1 1 1 1 1 1
<i>y</i> ₁	1 1 1 1	0 0 0 0 0 1
<i>y</i> ₃	1 0 1 1	0 0 0 0 0 0
<i>y</i> ₄	1 1 1 1	0 0 0 0 0 0
<i>y</i> ₅	0 1 1 1	1 1 0 0 0 0
<i>y</i> ₁₀	1 1 1 1	0 0 1 0 0 1
<i>y</i> ₁₁	1 1 1 1	0 0 0 0 0 0
<i>y</i> ₁₂	1 1 0 1	0 0 0 0 0 0
<i>y</i> ₁₃	1 1 1 1	0 0 0 0 0 1
<i>y</i> ₁₄	0 1 1 1	0 0 0 1 0 0
<i>y</i> ₁₅	0 1 0 1	0 0 0 0 1 0
<i>y</i> ₁₆	1 1 1 1	1 1 0 0 0 0
<i>y</i> ₁₇	1 1 1 1	0 0 0 1 0 1
<i>y</i> ₁₈	1 1 1 0	0 0 0 0 0 0
<i>y</i> ₁₉	1 0 1 1	1 0 1 0 0 0
<i>y</i> ₂₀	1 0 1 1	0 1 0 0 0 0

Matrice réorganisée

mode

0	1
1	0

Résumé

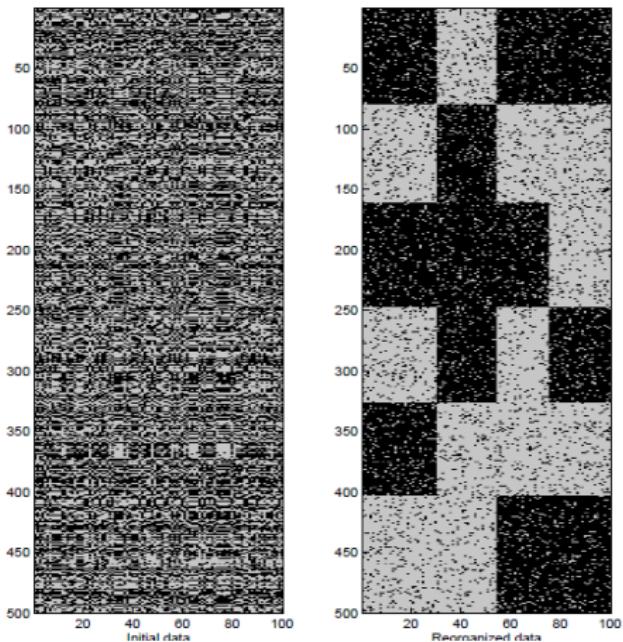
0.86	0.79
0.83	0.86

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 θ , 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):

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

- 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 $\ell(\theta; \mathbf{x})$ on θ is equivalent to maximize $\tilde{\ell}_c(\theta; \mathbf{x}, \mathbf{e})$ on (θ, \mathbf{e})

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

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 $\ell(\theta; \mathbf{x})$ or $\tilde{\ell}_c(\theta; \mathbf{x}, \mathbf{e})$ 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

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

MLE: variational EM (2/2)

Approximated E-step

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

- 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

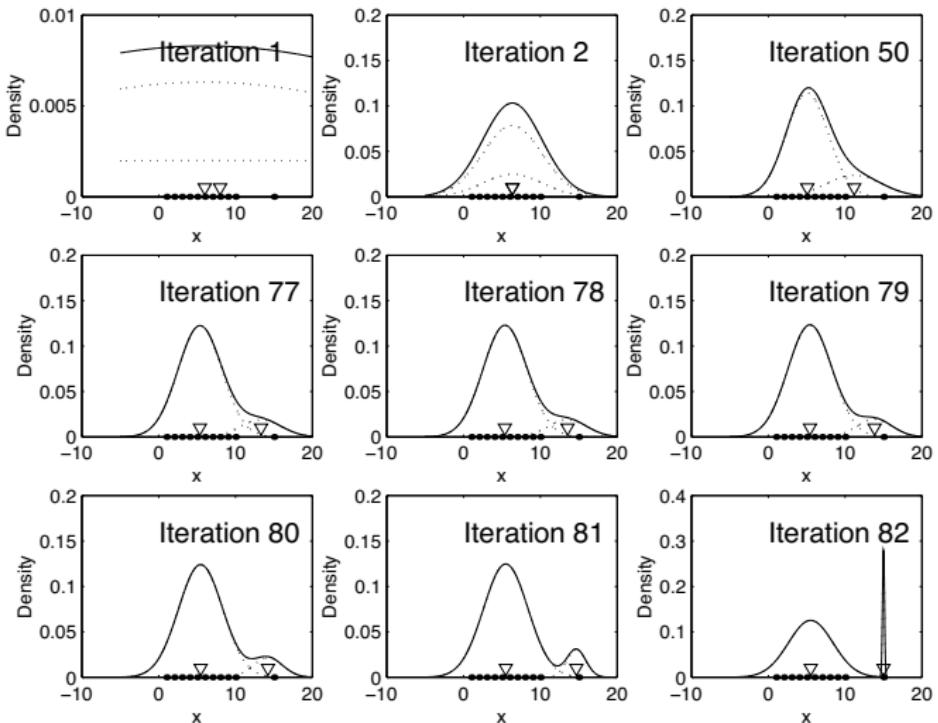
$$\mathbf{w}^{(q)} \sim p(\mathbf{w} | \mathbf{z}^{(q)}, \mathbf{x}; \boldsymbol{\theta}^{(q)})$$

- 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 θ**

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

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

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

Bayesian estimation: limiting degeneracy

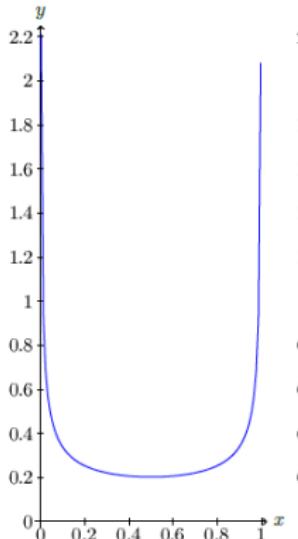
- Interest for avoiding degeneracy is the prior: it acts as a **penalization** term
- Typical choices are **Dirichlet** for π and ρ (with independence between π , ρ , α)

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

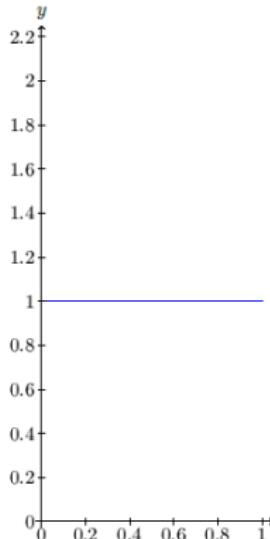
- The Dirichlet distribution is conjugate, thus easy calculus
- Control degeneracy frequency with the a value:**
 - $a = 1$: uniform prior, so $\hat{\theta}$ is strictly the MLE (no regularisation)
 - $a = 1/2$: Jeffreys prior, classical (no informative prior) but may favor degeneracy
 - $a = 4$: a rule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

Bayesian estimation: prior overview

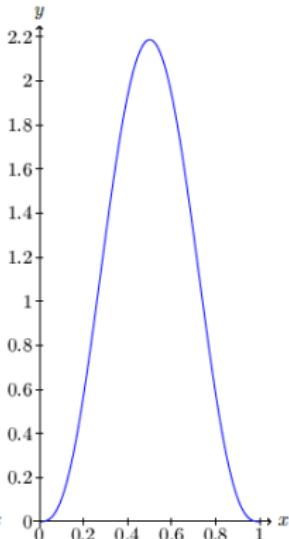
$$\text{Be}\left(\frac{1}{2}, \frac{1}{2}\right)$$



$$\text{Be}(1, 1)$$

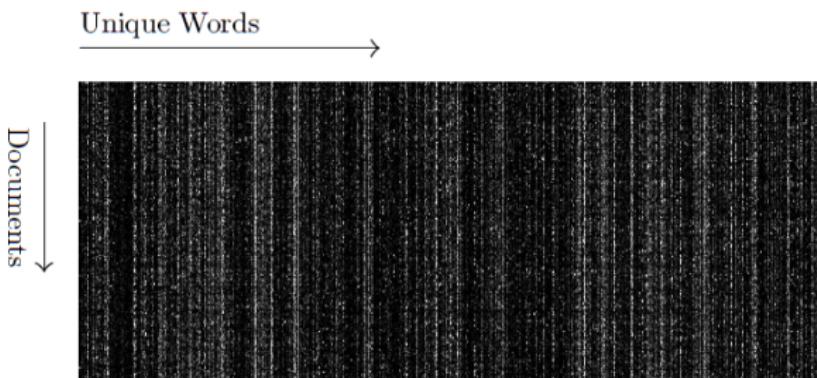


$$\text{Be}(4, 4)$$



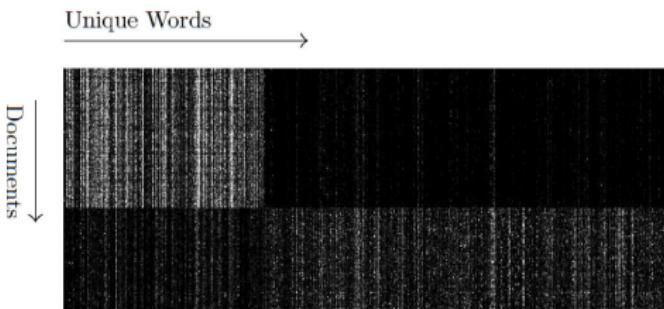
Document clustering (1/2)

- 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×words
- Poisson model



Document clustering (2/2)

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



Results with 2x2 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	Trial BlockCluster
Data File	Blockcluster-Example.csv
Data Type	Categorical
Rows Cluster Groups	1:5
Column Cluster Groups	1:5
<button>Update</button>	

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	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

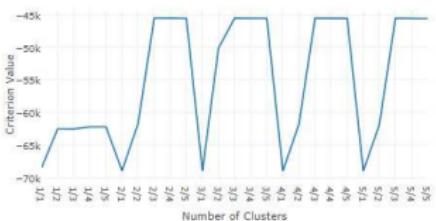
Model	Criterion	Nb Clusters	Error
<i>pik_rho_multi</i>	ICL (-45557.1)	[2,3]	No error
<i>pik_rho_multi</i>	ICL (-45563.3)	[3,3]	No error
<i>pik_rho_multi</i>	ICL (-45566.6)	[2,4]	No error
<i>pik_rho_multi</i>	ICL (-45573.9)	[4,3]	No error
<i>pik_rho_multi</i>	ICL (-45574.6)	[5,3]	No error
<i>pik_rho_multi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

Cluster Plot Criterion Plot

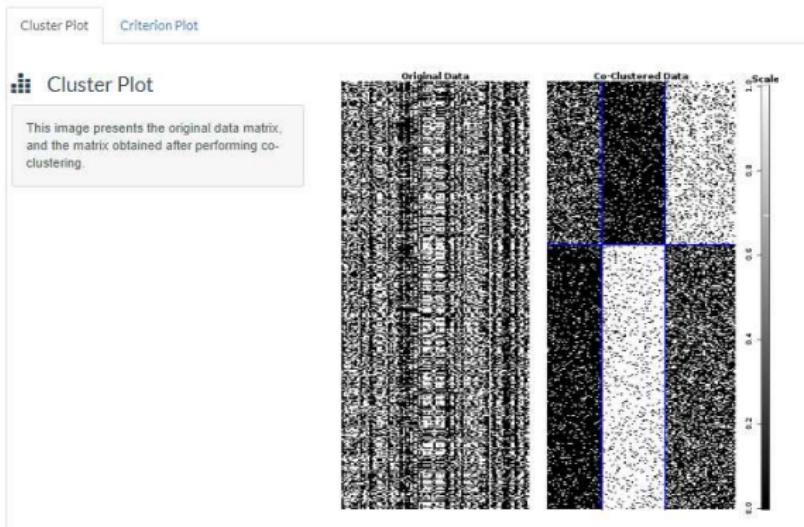
Model Criterion

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

ICL value / Nb of Clusters



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/>