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

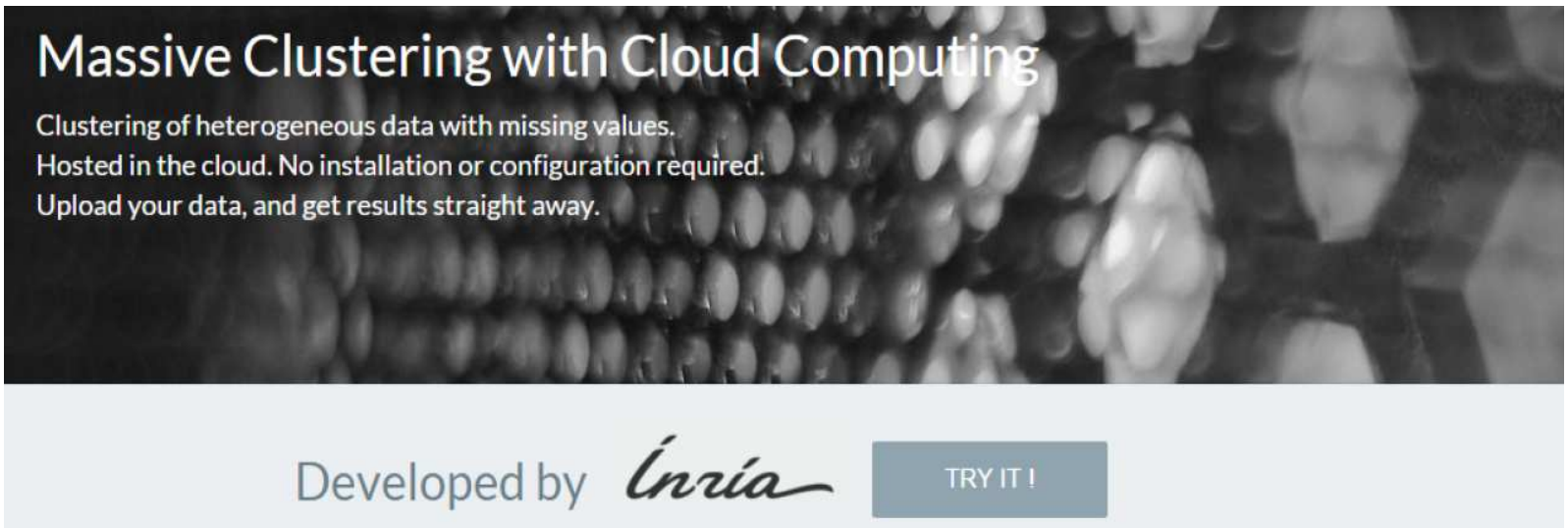
21-22 June 2018, Workshop MixStatSeq, Paris

"Mixture models: Theory and applications"

Outline

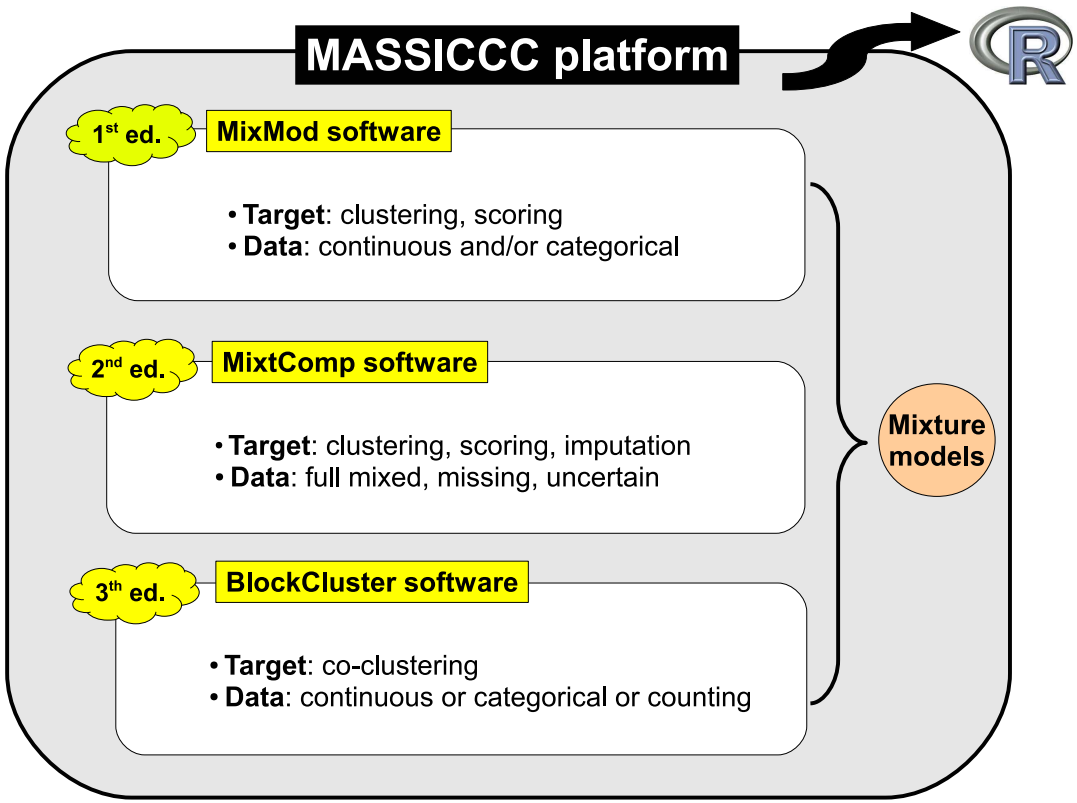
MASSICCC?

massiccc.lille.inria.fr



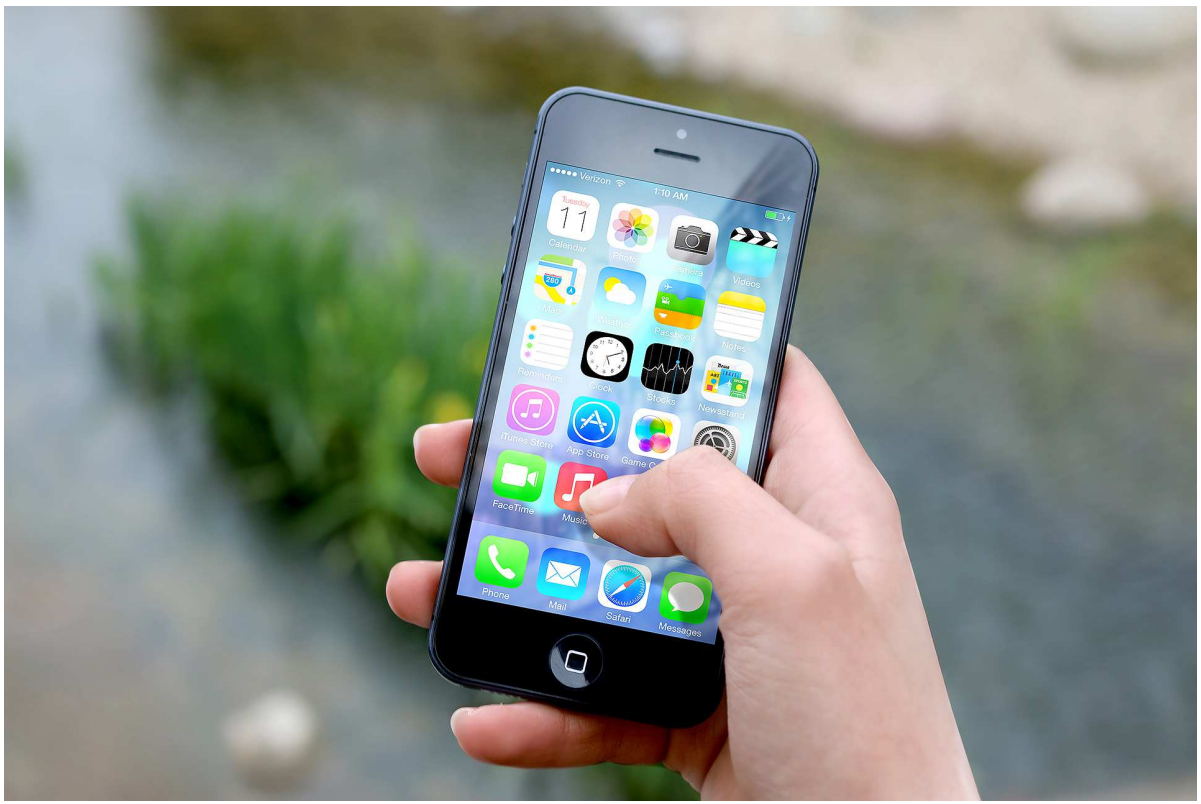
SaaS: Software as a Service

MASSICCC??



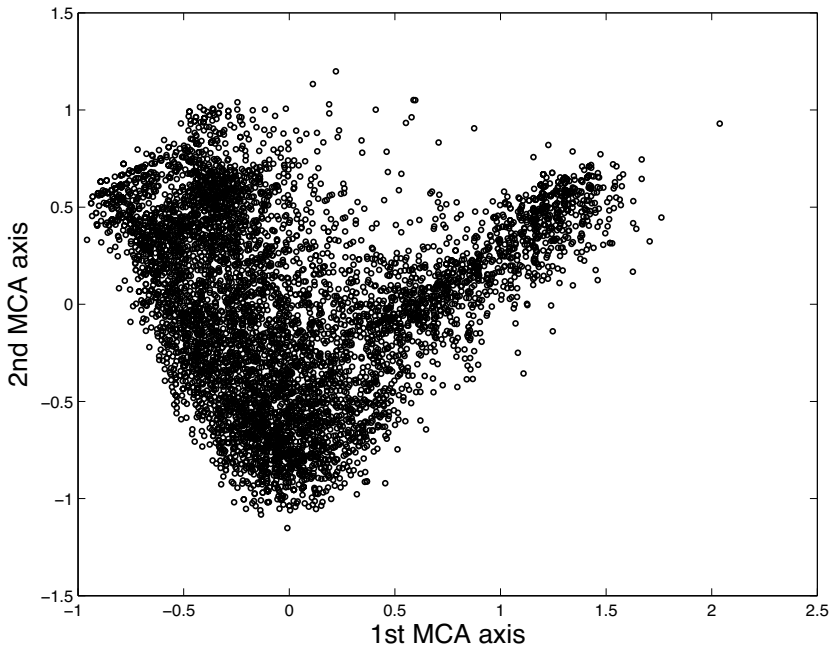
A high quality and easy to use web platform
where are transfered mature research clustering (and more) software
towards (non academic) professionals

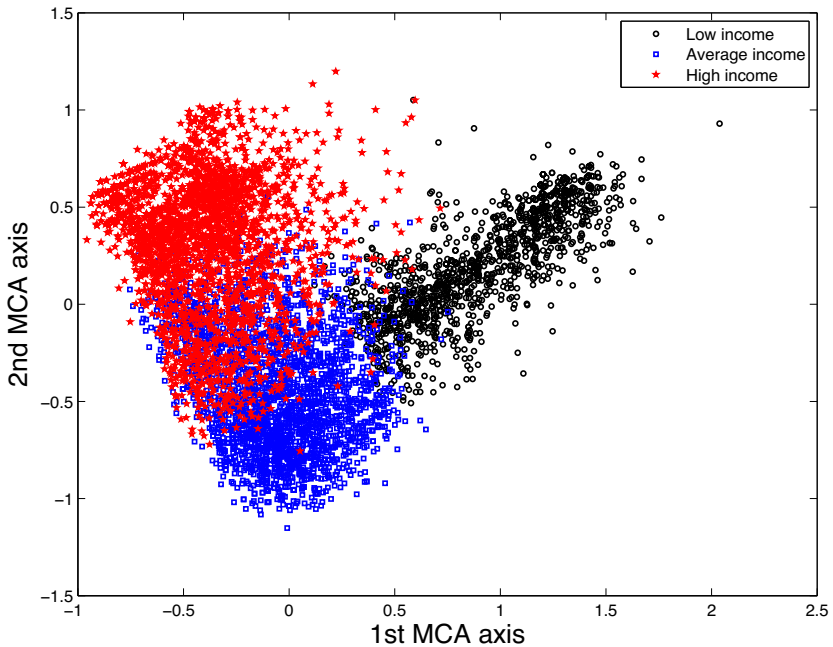
Here is the computer you need!



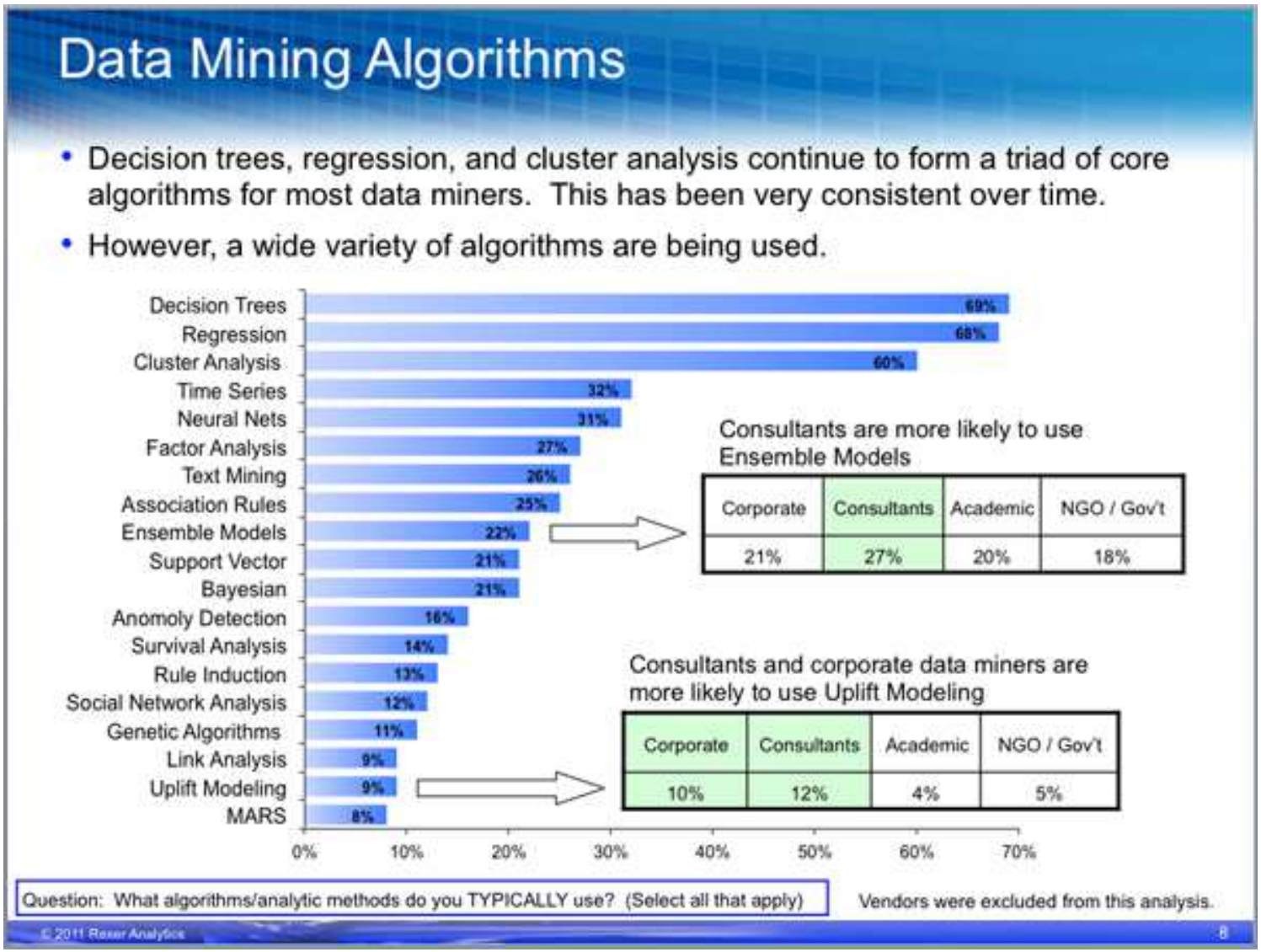
Clustering?

Detect hidden structures in data sets



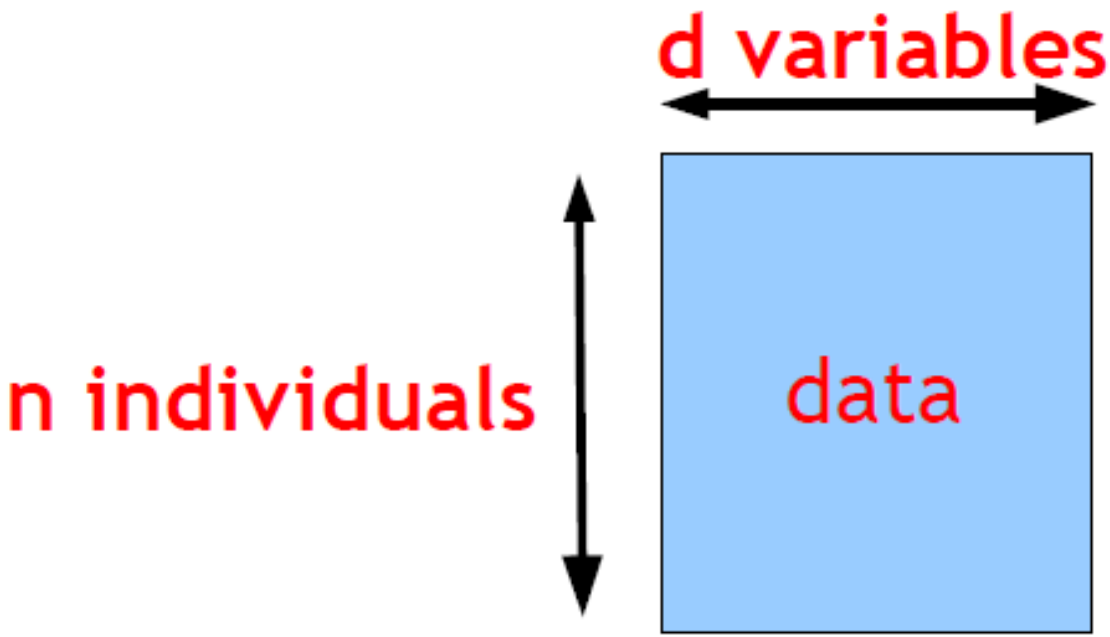


Clustering everywhere[1]

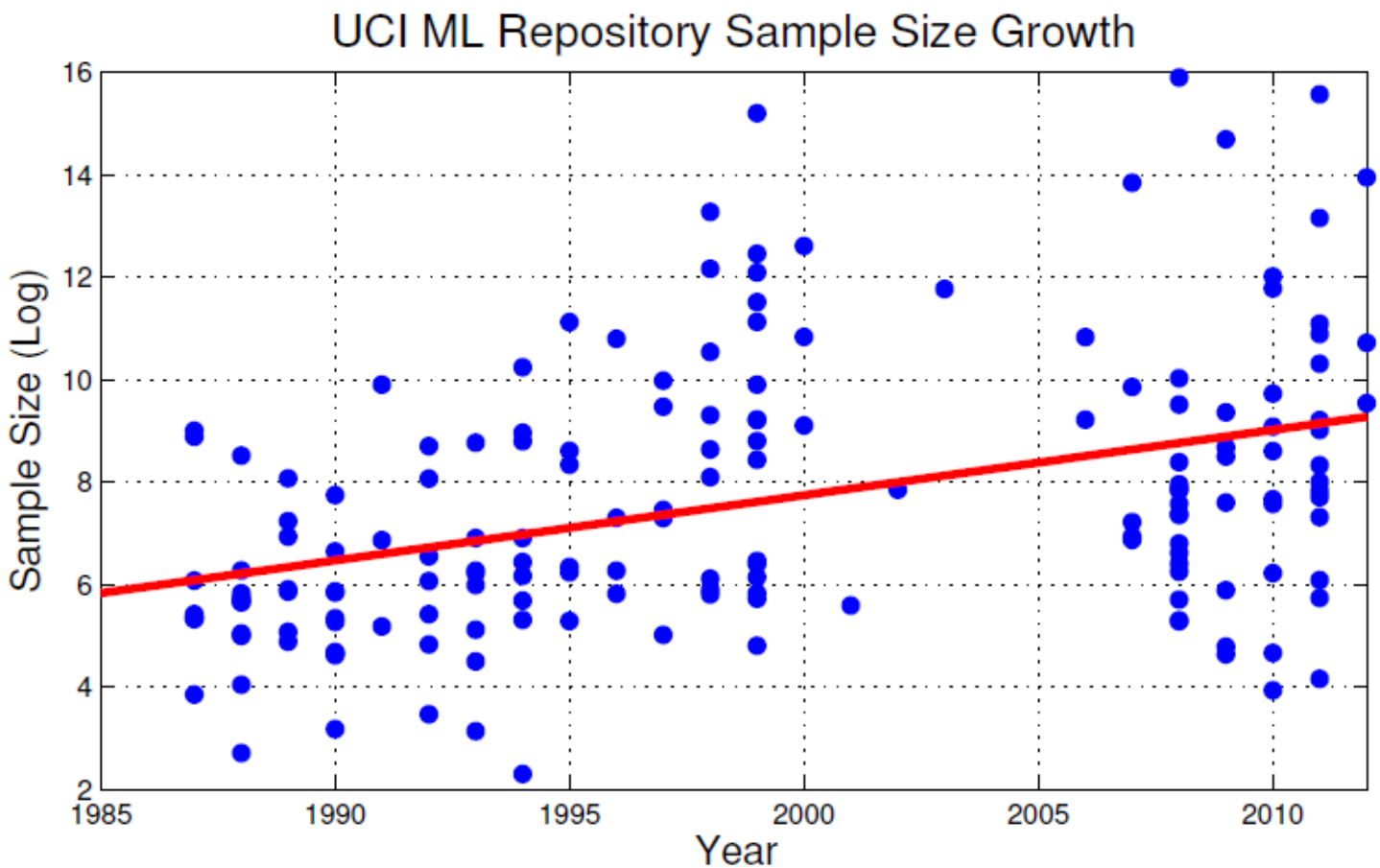


[1] Rexer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

Data sets structure

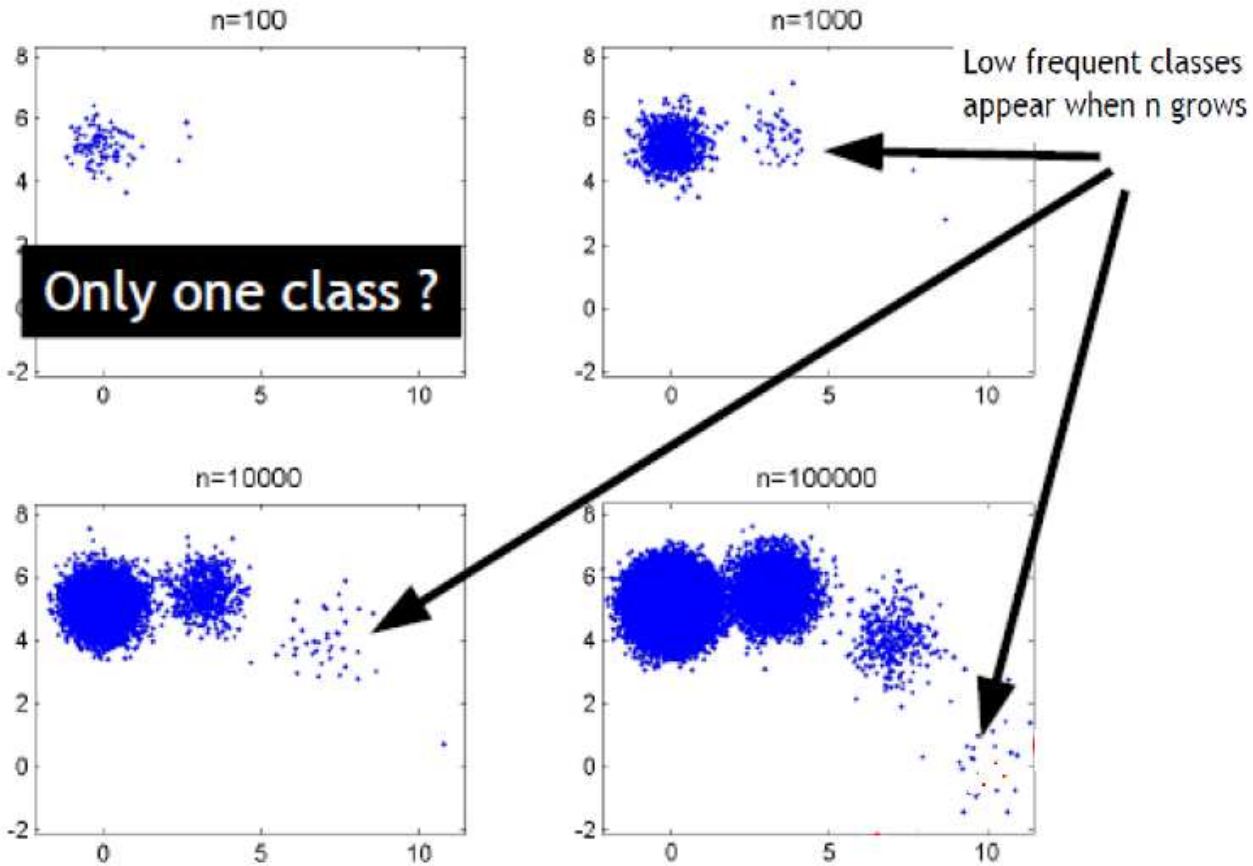


Large data sets[2]



[2]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



Todays features: full mixed/missing

categorical
Marital status
maried

integer
Children
3

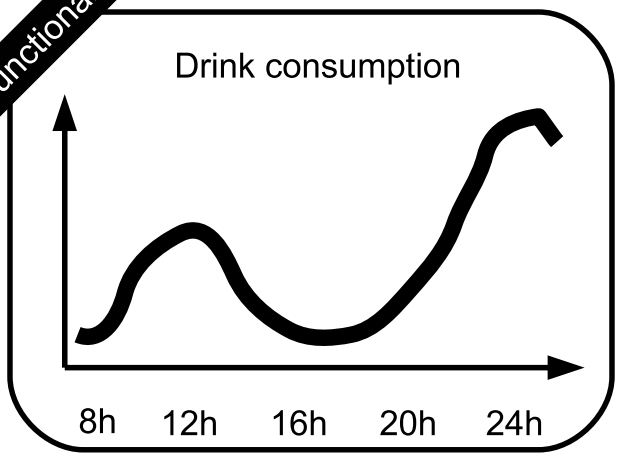
missing
Size (m)
?

rank
Drink preference
beer > soda > water

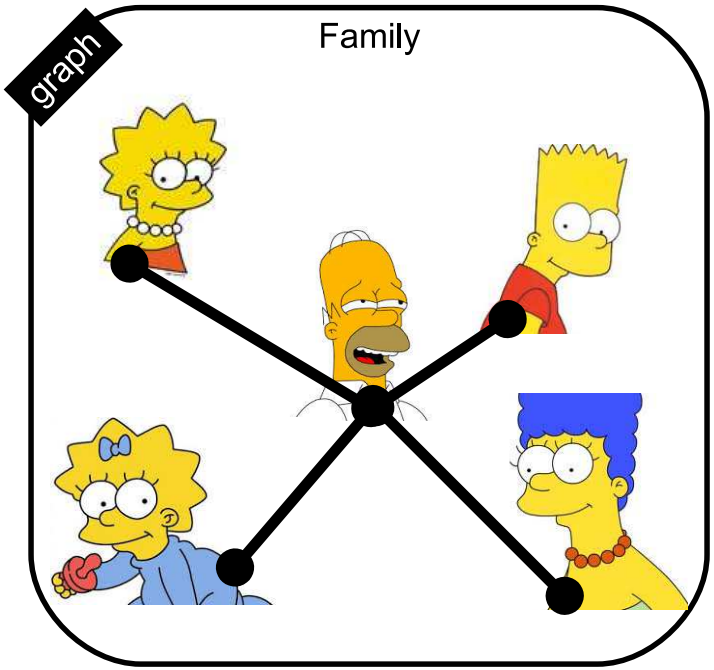
ordinal
Intelligence
low

continuous
Weight (kg)
119.5

functional



graph



And so on...

Notations

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

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

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

Parametric mixture model

- Parametric assumption:

$$p_k(\mathbf{x}_1) = p(\mathbf{x}_1; \boldsymbol{\alpha}_k)$$

thus

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

- Mixture parameter:

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha}) \text{ with } \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$$

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

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

The number of free *continuous* parameters is given by

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

Clustering becomes a well-posed problem...

The clustering process in mixtures

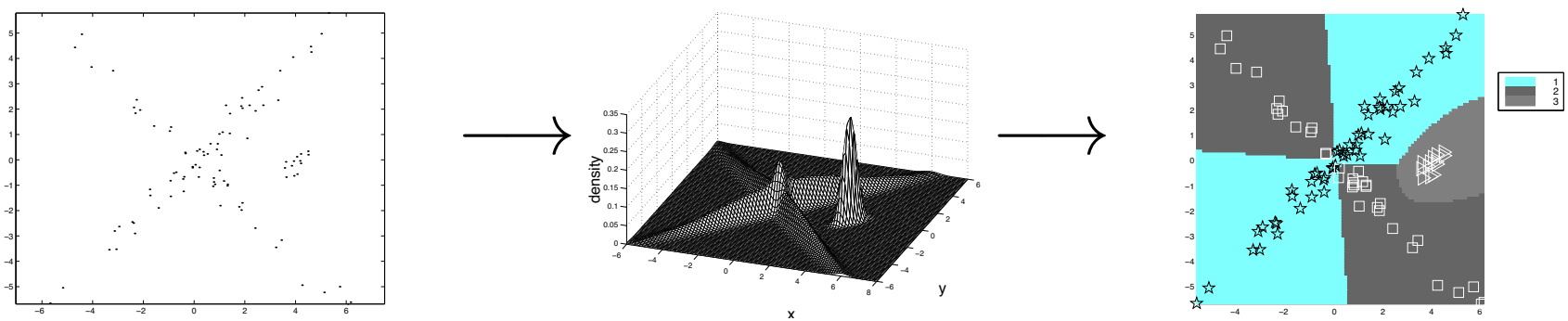
1. Estimation of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$
2. Estimation of the conditional probability that $\mathbf{x}_i \in G_k$

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

3. Estimation of $\mathbf{z}_i$ by *maximum a posteriori* (MAP)

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

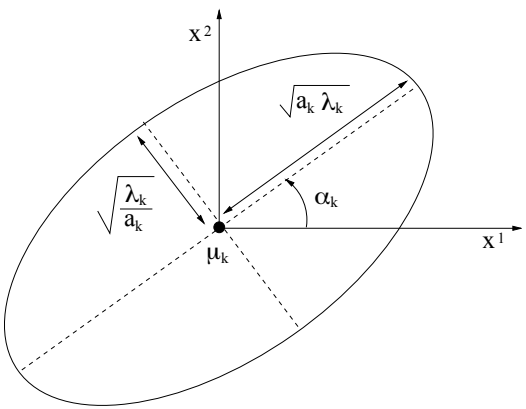
4. Model selection: BIC, ICL, ...

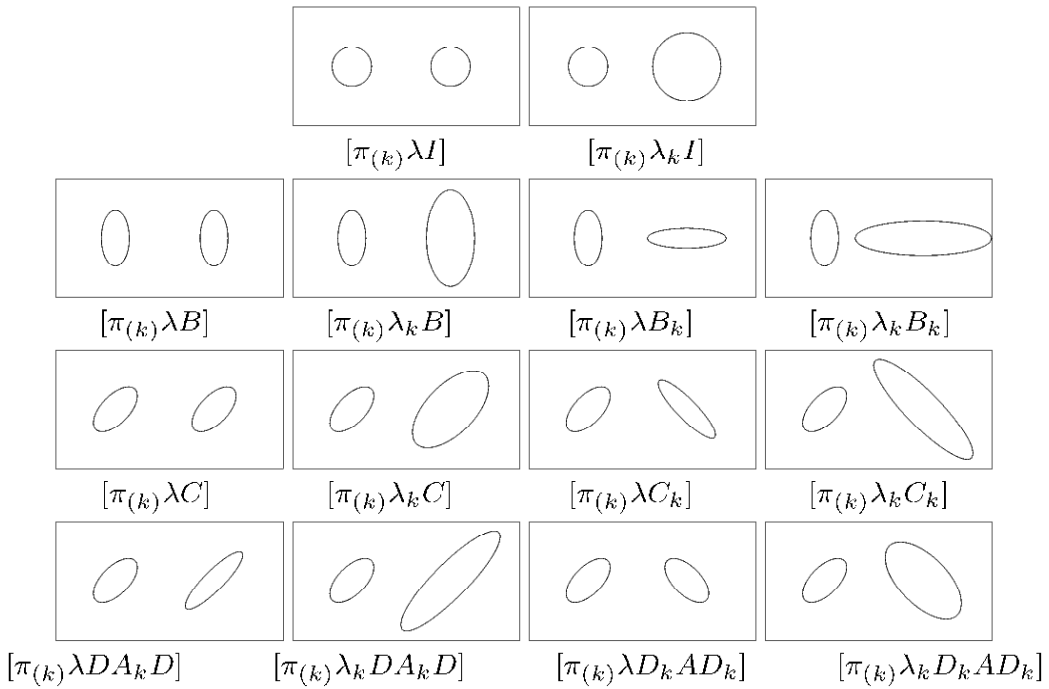


Outline

Only continuous features: 14 models on $\mathbf{\Sigma}_k$

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





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:

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

and

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

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

Mixing continuous and categorical data: full local independence

Combine continuous and categorical data

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

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

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

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 $\boldsymbol{\theta}$ by *complete*-likelihood

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

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

- 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 $\hat{\boldsymbol{\theta}}$: 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 $\boldsymbol{\theta}$ by *observe*-likelihood

Maximize the observe-likelihood on $\boldsymbol{\theta}$

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

- Convergence of $\hat{\boldsymbol{\theta}}$, 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

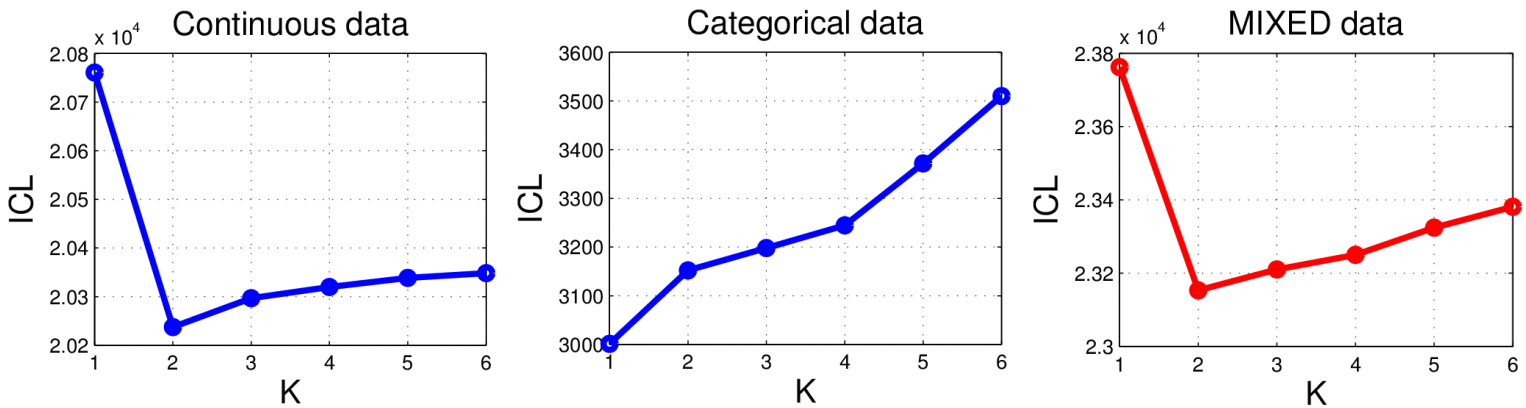
- Initialization: $\boldsymbol{\theta}^0$
- Iteration n^o q:
  - Step E: estimate probabilities $\mathbf{t}^q = \{t_{ik}(\boldsymbol{\theta}^q)\}$
  - Step C: classify by setting $\mathbf{t}^q = \text{MAP}(\{t_{ik}(\boldsymbol{\theta}^q)\})$
  - Step M: maximize $\boldsymbol{\theta}^{q+1} = \arg\max_{\boldsymbol{\theta}}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

Properties

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

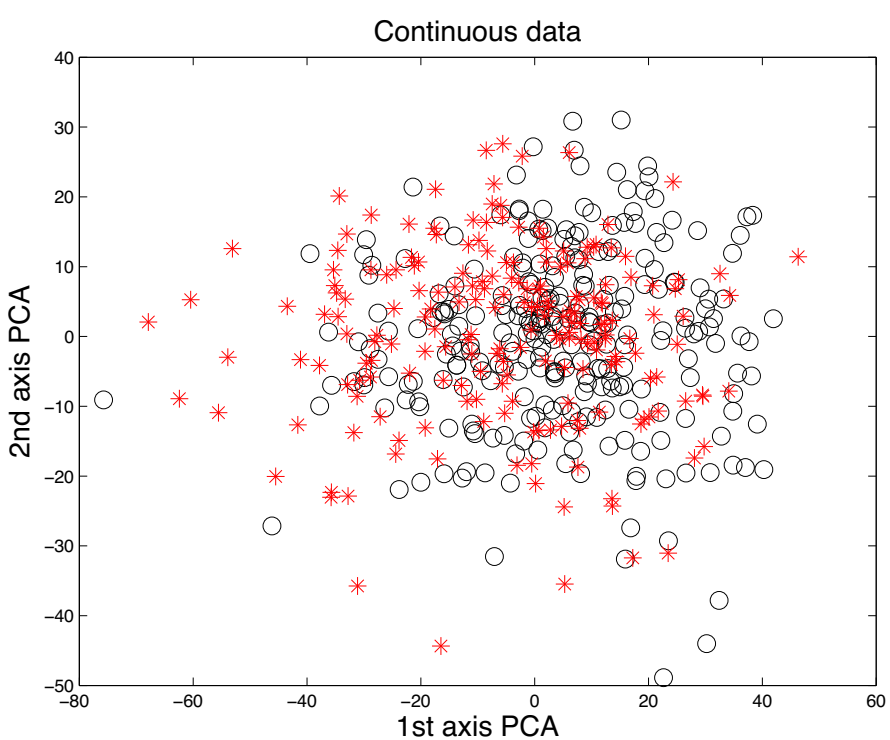
Prostate cancer data (without mixing data)

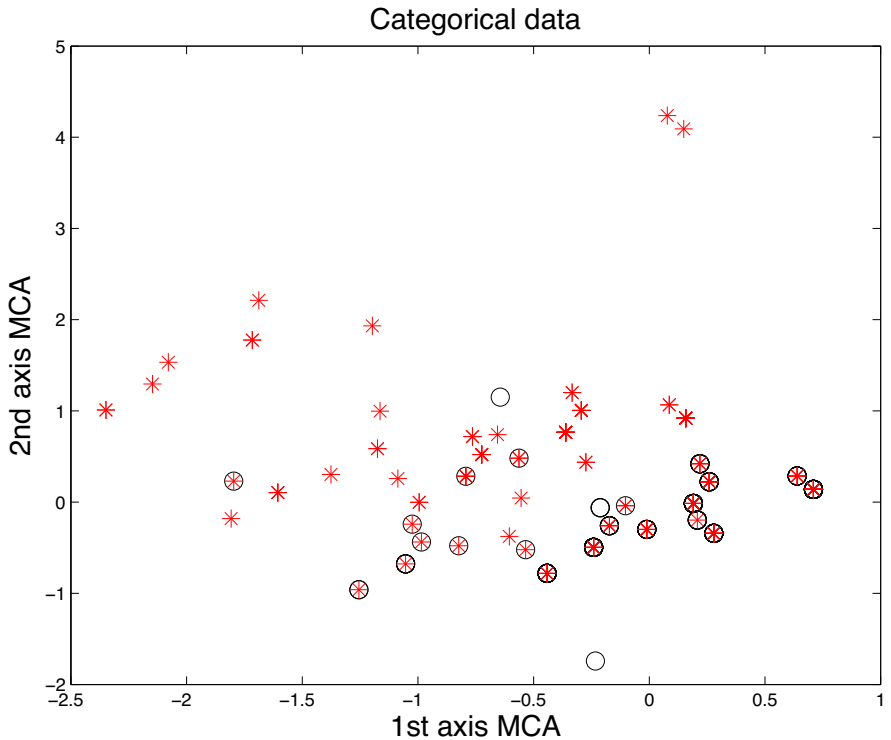
- Individuals: $n = 475$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic 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. $\mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k) = \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cont}) \cdot \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cat})$



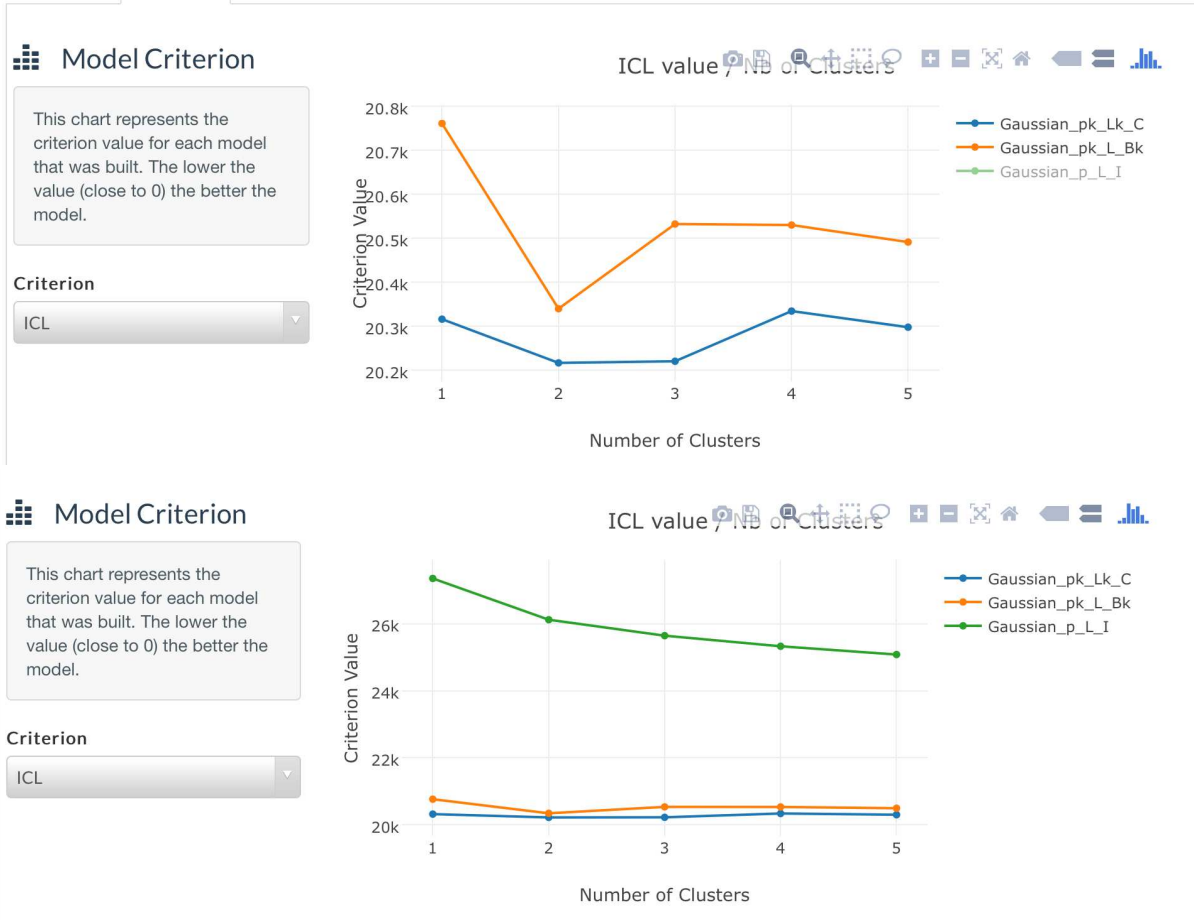
Prostate cancer data (without missing data)

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

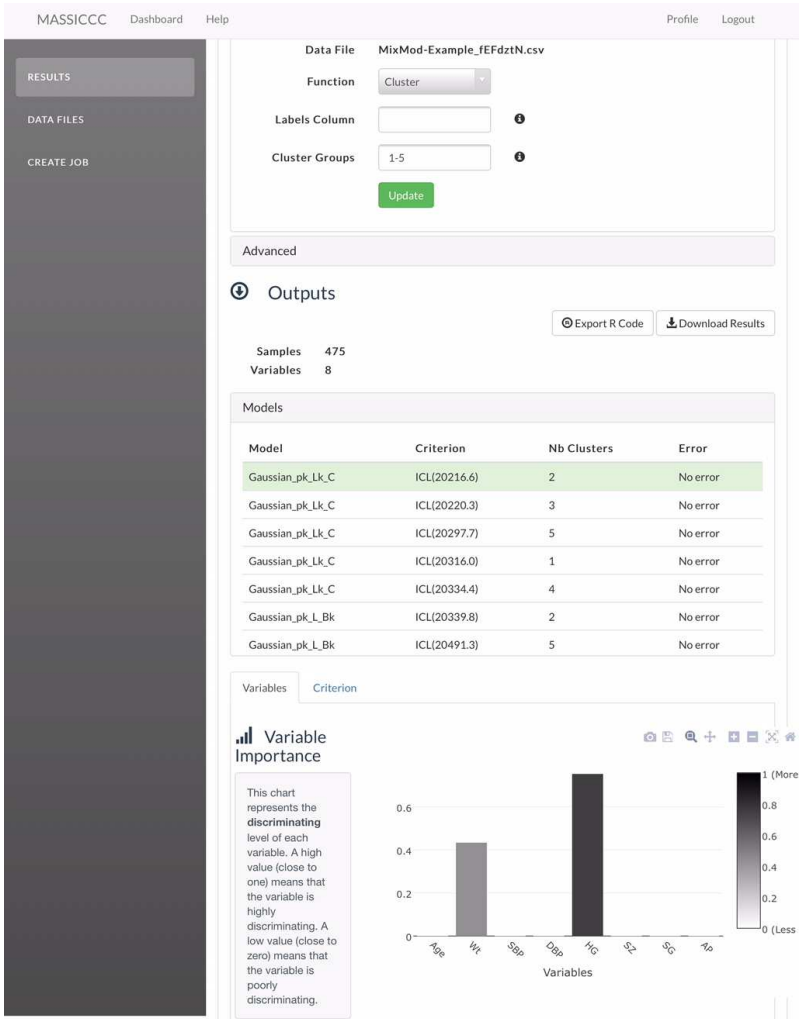


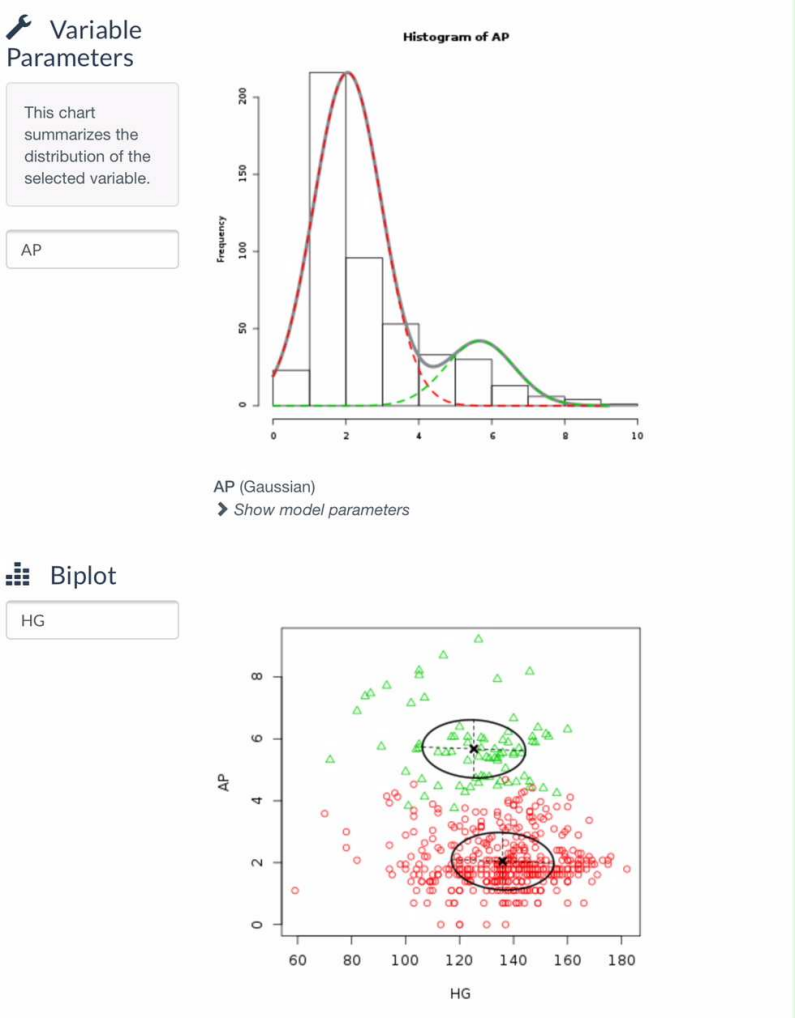


Continuous data

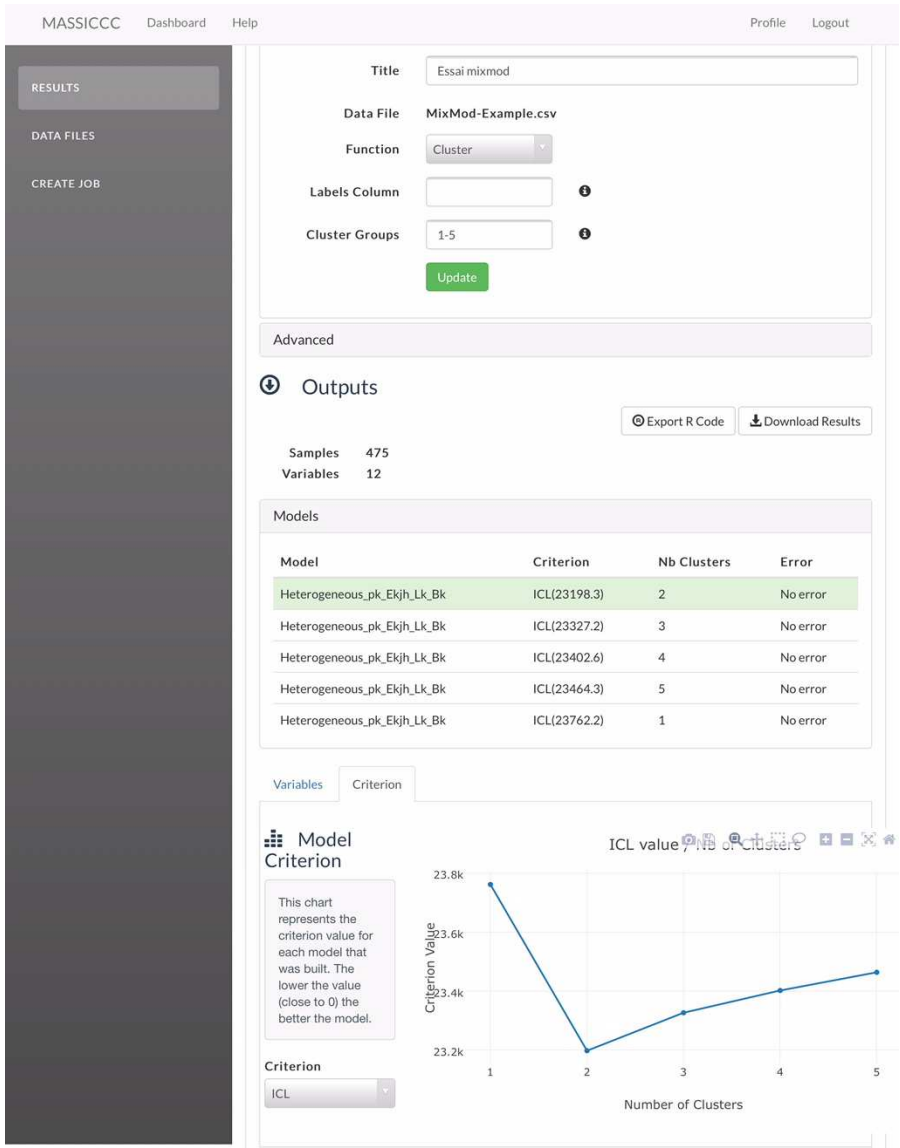


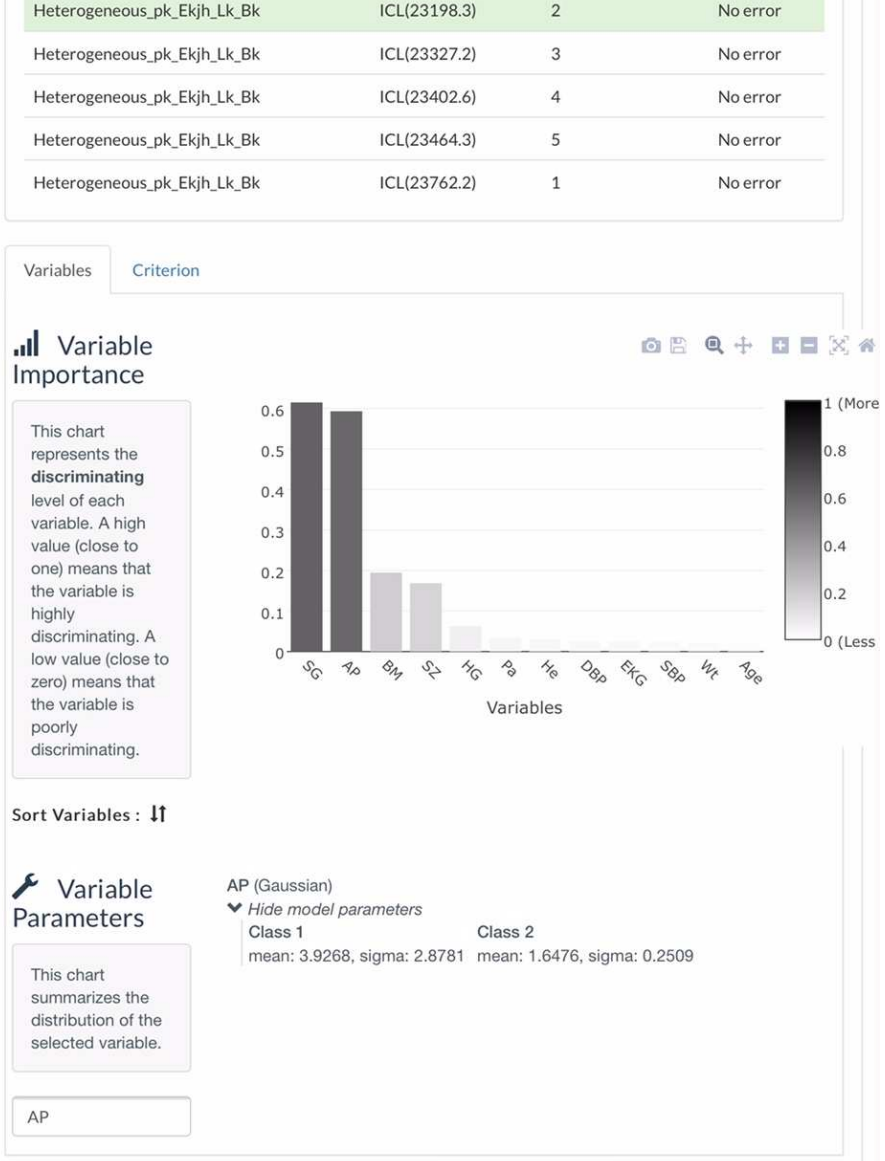
Continuous data





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_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
ICL value / Nb of clusters
Criterion Value
23.8k
23.6k
23.4k
23.2k
1
2
3
4
5
Number of Clusters


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.
0.6
0.5
0.4
0.3
0.2
0.1
0
SG
AP
BM
SZ
HG
Pa
He
DBP
EKG
SBP
Wt
Age
Variables
1 (More
0.8
0.6
0.4
0.2
0 (Less
Sort Variables :
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

Full mixed data: conditional independence everywhere[3]

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

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

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

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

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
- ...

[3] MixtComp software on the MASSICCC platform: `https://massiccc.lille.inria.fr/`

Missing data: MAR assumption and estimation

Assumption on the missingness mecanism

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

Observed log-likelihood...

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

SEM algorithm[4]

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

- Initialisation: $\boldsymbol{\theta}^{(0)}$
- Iteration nb q:
  - E-step: compute conditional probabilities $\mathrm{p}(\boldsymbol{x}^M, \boldsymbol{z}|\boldsymbol{x}^0; \boldsymbol{\theta}^{(q)})$
  - S-step: draw $(\boldsymbol{x}^{M(q)}, \boldsymbol{z}^{(q)})$ from $\mathrm{p}(\boldsymbol{x}^M, \boldsymbol{z}|\boldsymbol{x}^0; \boldsymbol{\theta}^{(q)})$
  - M-step: maximize $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} \ln \mathrm{p}(\mathbf{x}^O, \boldsymbol{x}^{M(q)}, \boldsymbol{z}^{(q)}; \boldsymbol{\theta})$
- 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 $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}$
- The variance of the sequence $(\boldsymbol{\theta}^{(q)})$ gives confidence intervals

[4] MixtComp 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 histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Some missing data: 62 missing values ($\approx 1\%$)

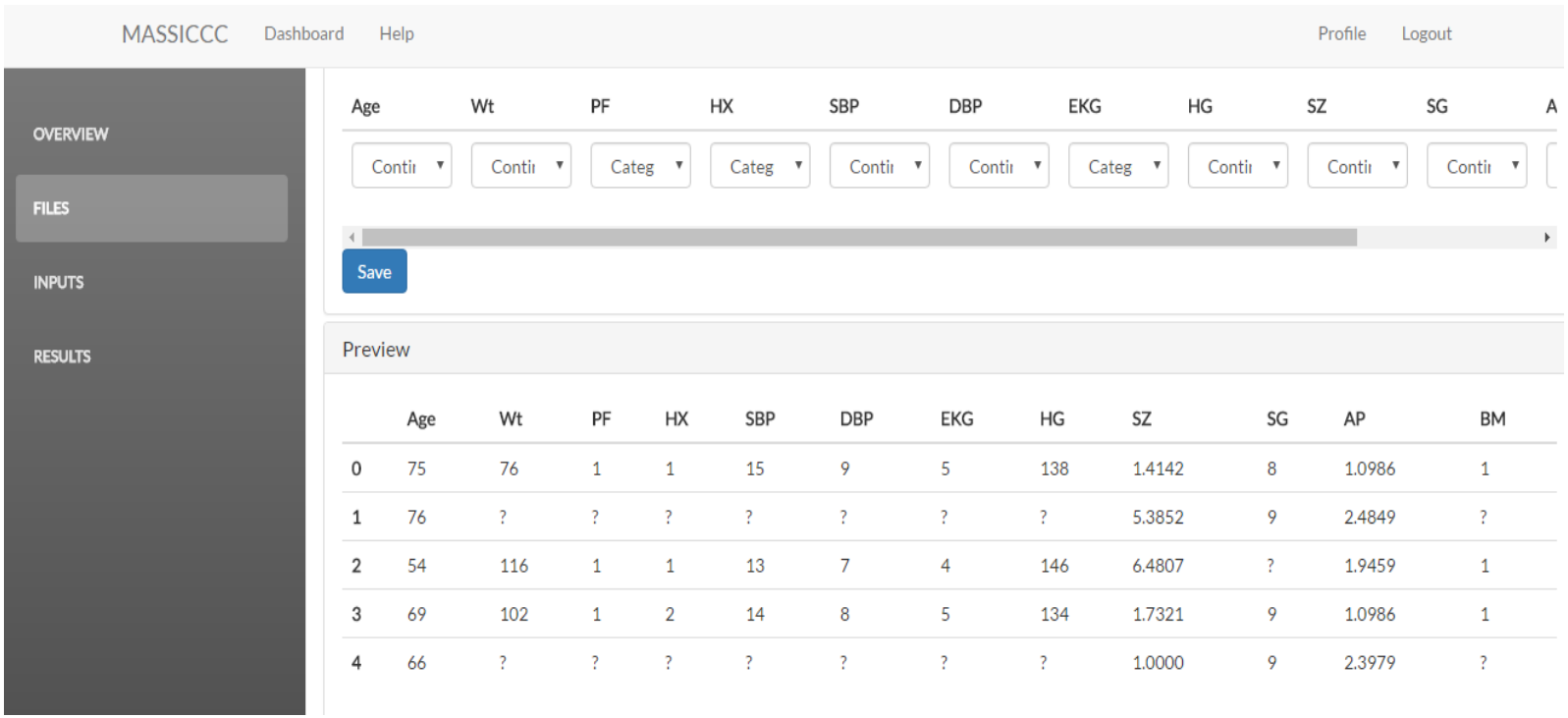
We forget the classes (Stages of the desease) for performing clustering

Questions

- How many clusters?
- Which partition?

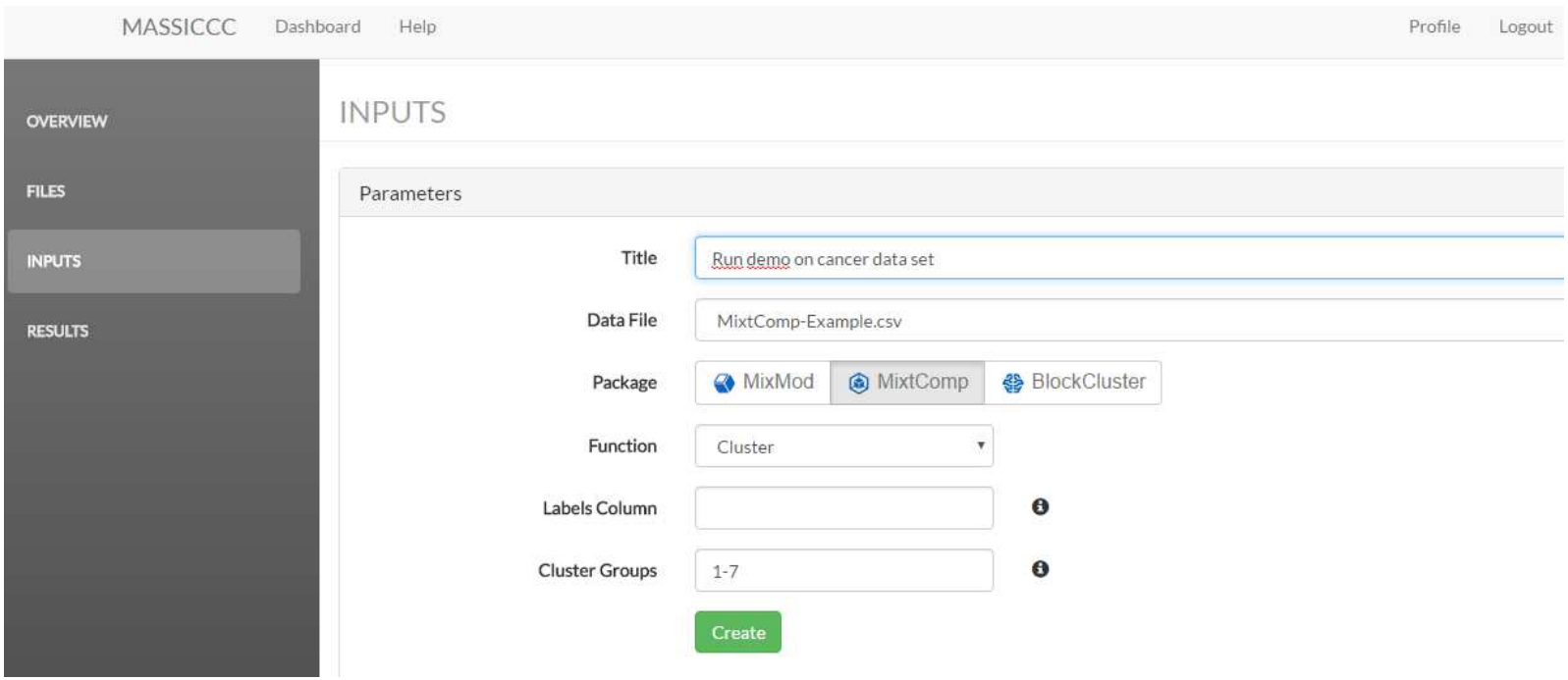
[5] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Data upload without preprocessing

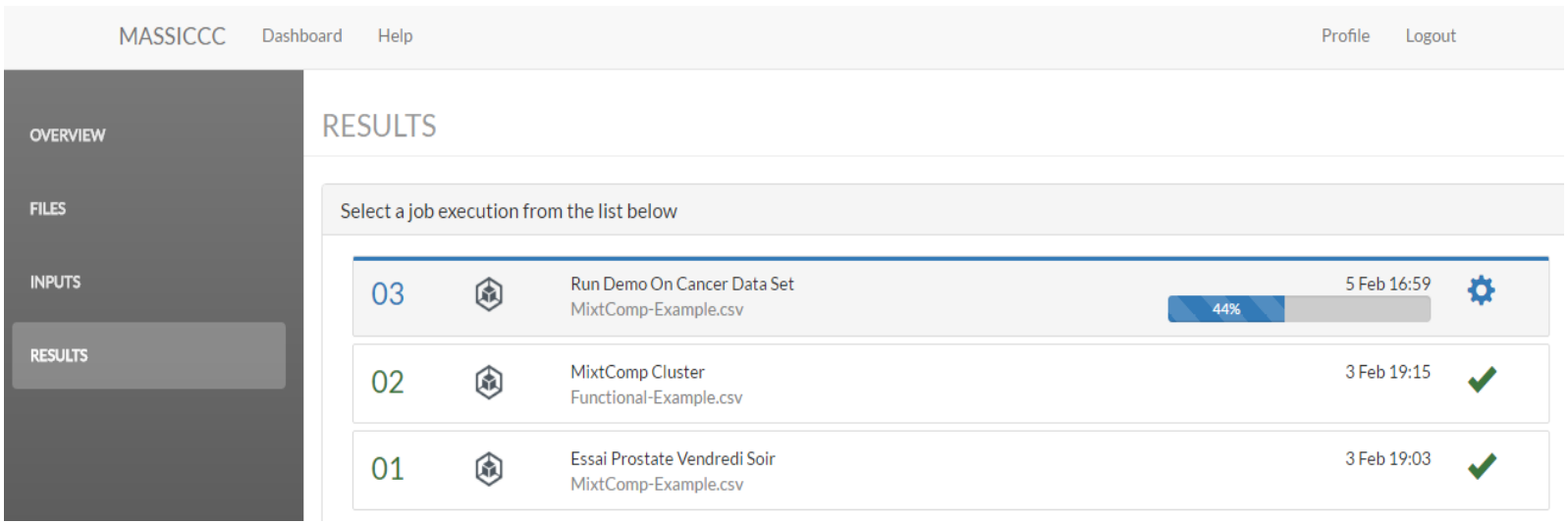


	Age	Wt	PF	HX	SBP	DBP	EKG	HG	SZ	SG	AP	BM
0	75	76	1	1	15	9	5	138	1.4142	8	1.0986	1
1	76	?	?	?	?	?	?	?	5.3852	9	2.4849	?
2	54	116	1	1	13	7	4	146	6.4807	?	1.9459	1
3	69	102	1	2	14	8	5	134	1.7321	9	1.0986	1
4	66	?	?	?	?	?	?	?	1.0000	9	2.3979	?

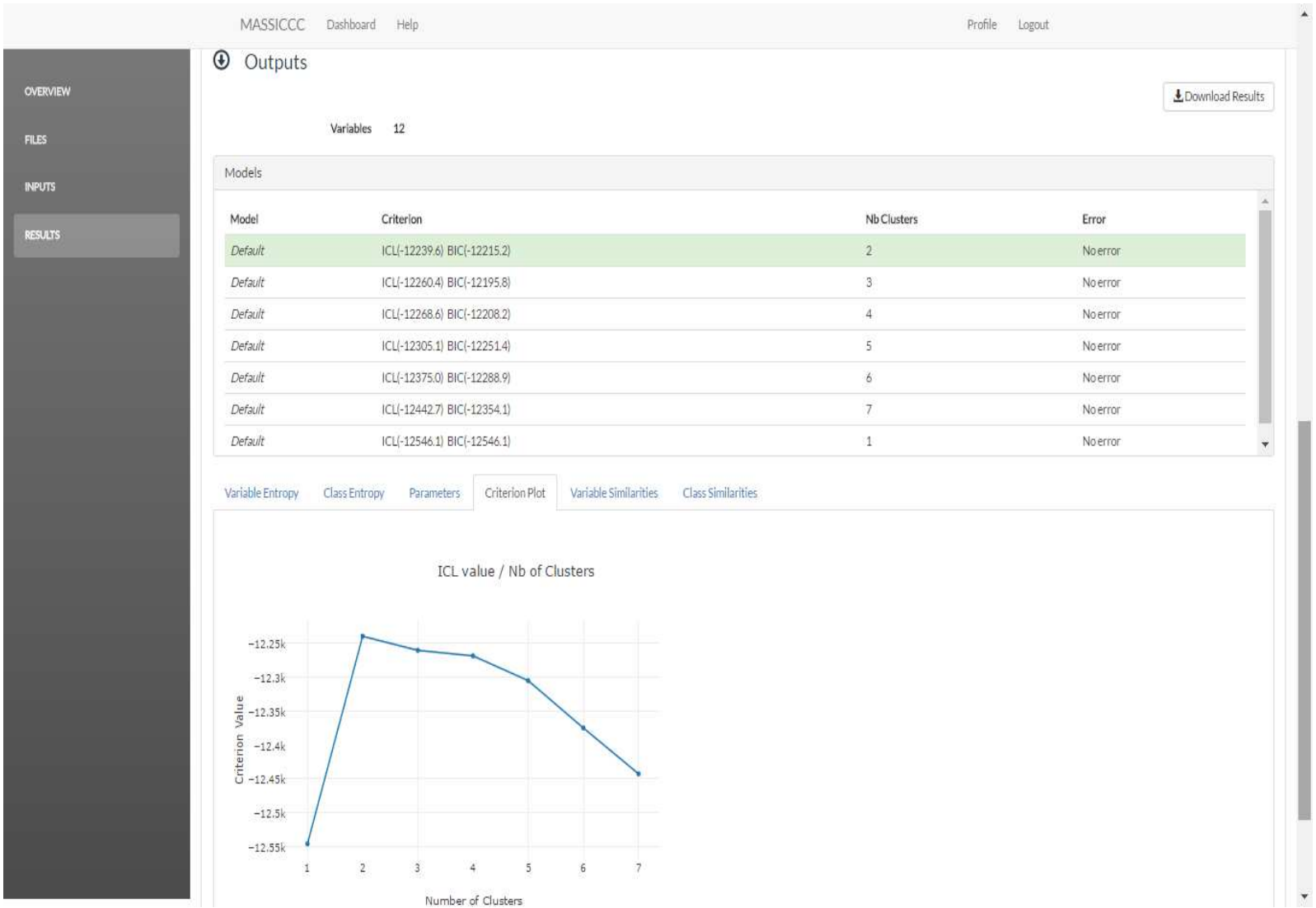
Run clustering analysis



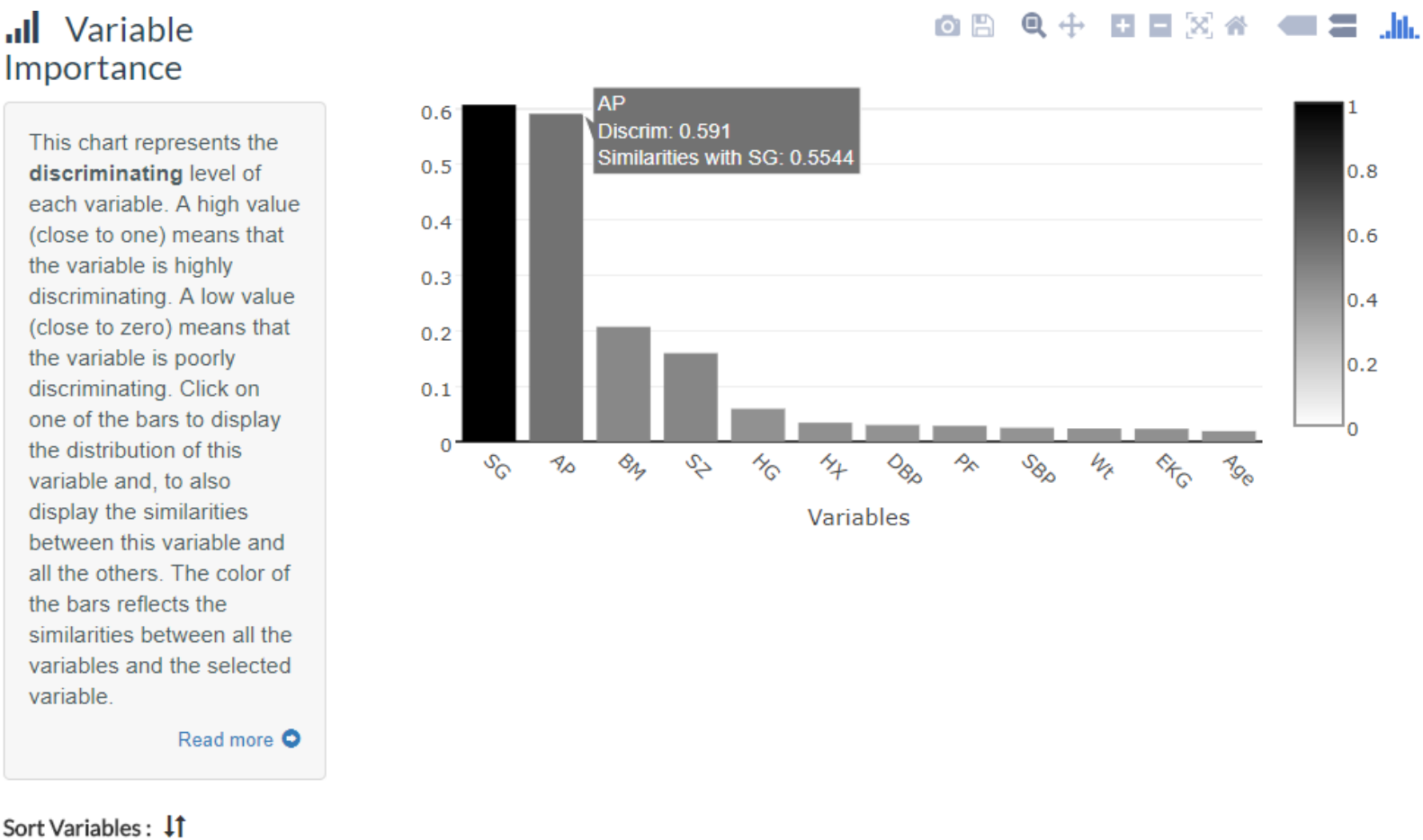
It is running on the (Inria) cloud...



Several quick result overviews. . . without post-processing



Variable significance on global partition



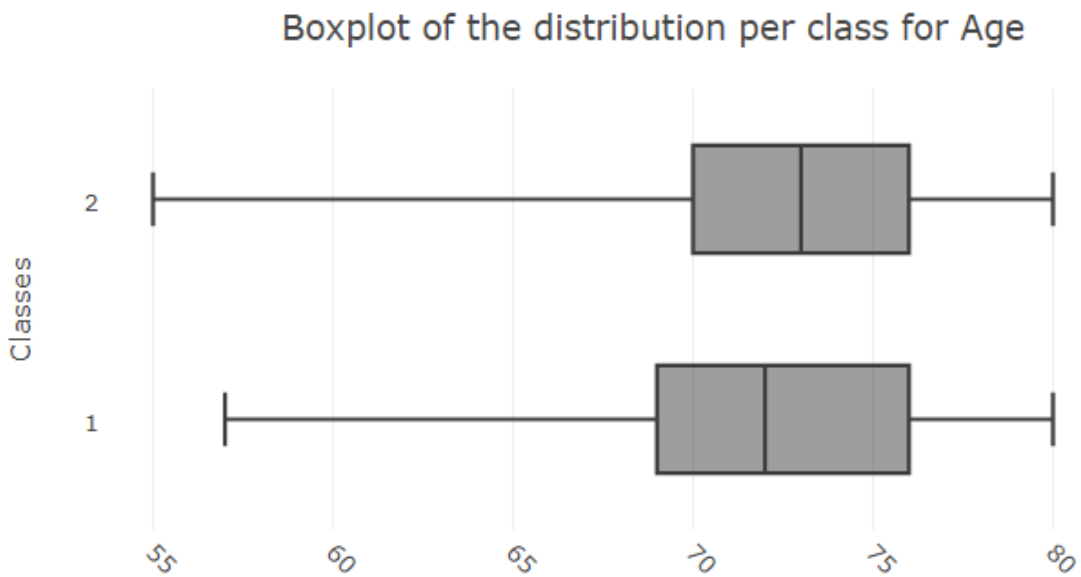
+ similarity between variables

Variable "Age" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

Age



Age (Gaussian)

Hide model parameters

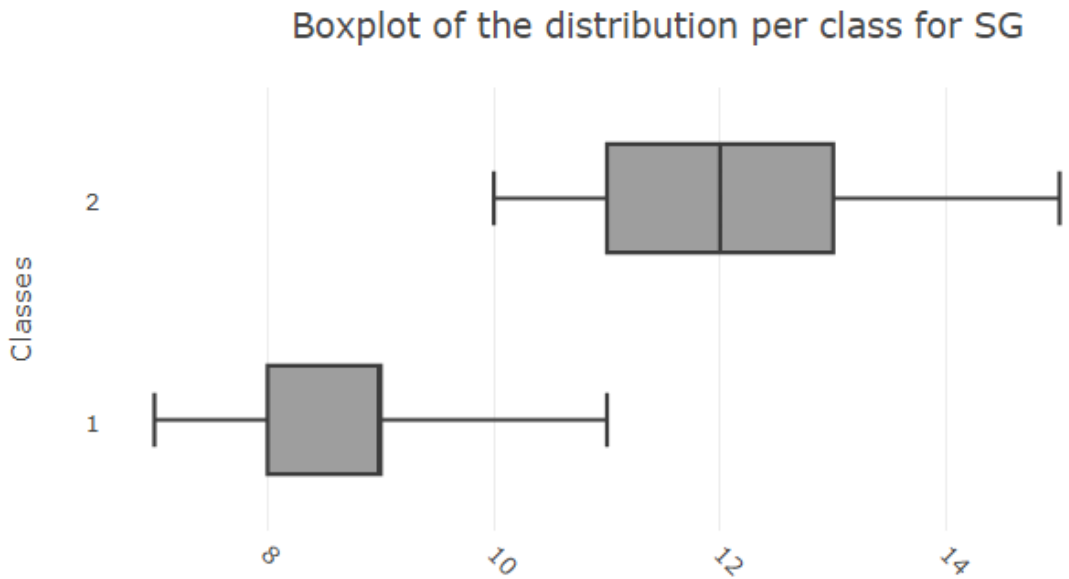
Class 1	Class 2
mean: 71.534, sigma: 6.760	mean: 71.313, sigma: 7.463

Variable "SG" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

SG



SG (Gaussian)

Hide model parameters

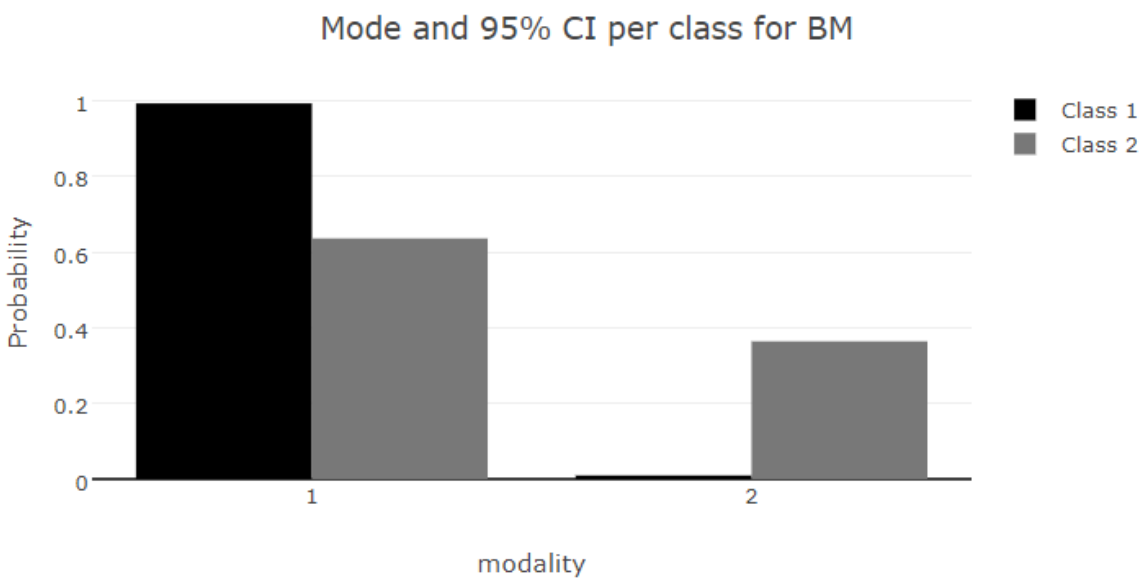
Class 1	Class 2
mean: 8.940, sigma: 1.154	mean: 12.087, sigma: 1.405

Variable "BM" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

BM

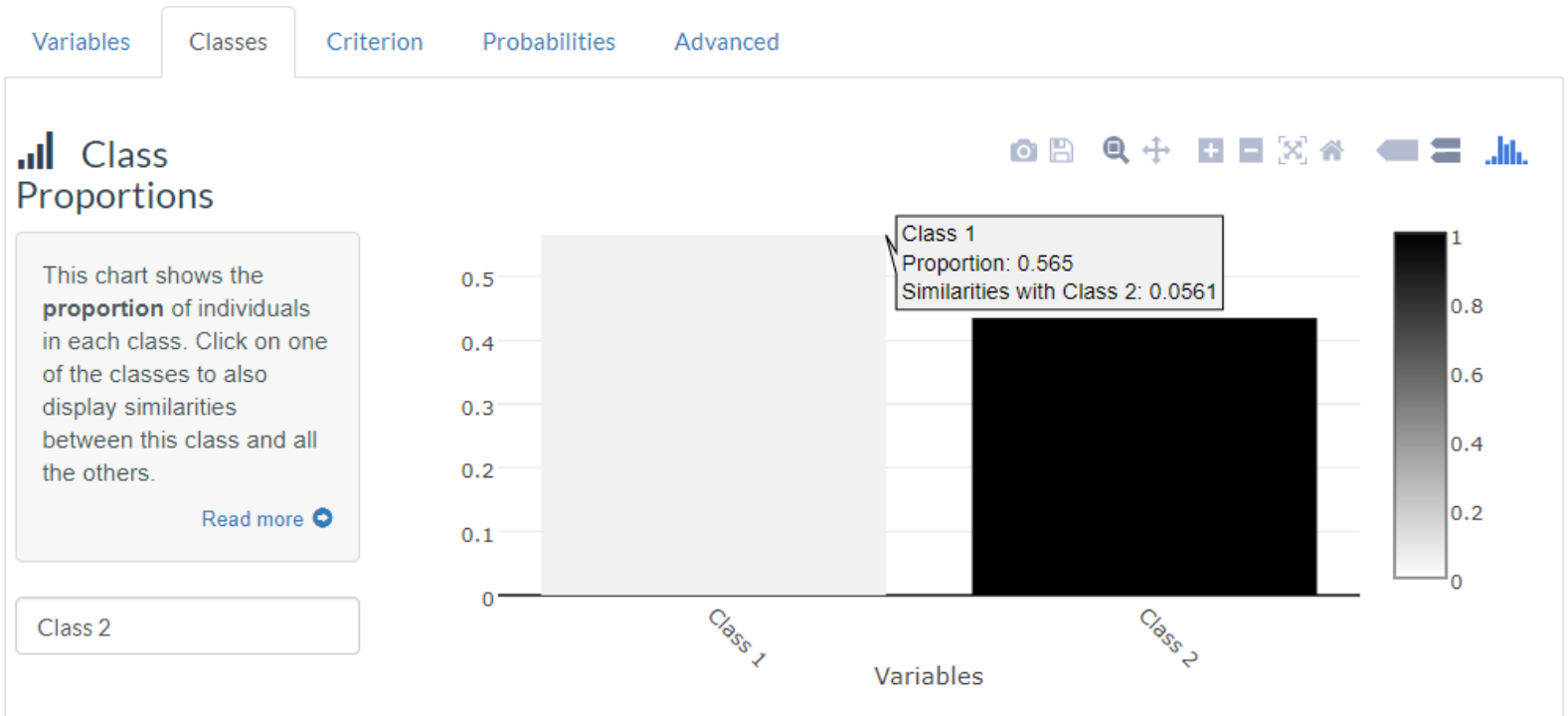


BM (Multinomial)

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[6] + MixtComp"**: MixtComp on the dataset completed by `mice`

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

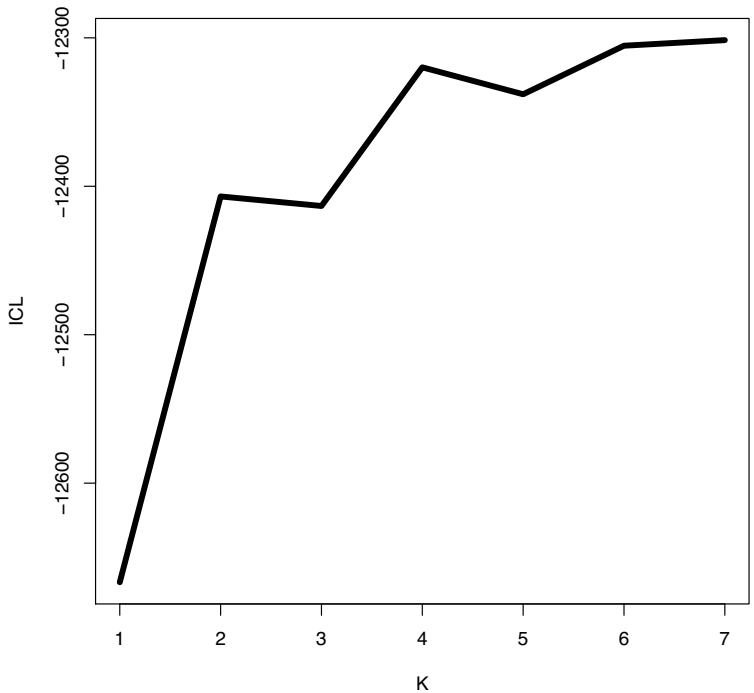
- **Strategy "full MixtComp"**: MixtComp on the observed (no completed) dataset

Partition quality with $K = 2$

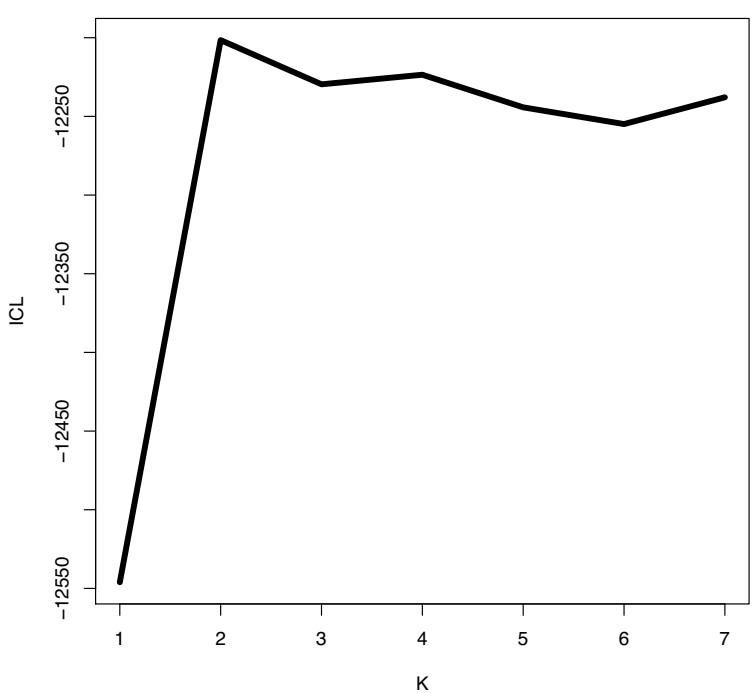
Strategy	mice + MixtComp	full MixtComp
% misclassified	12.8	8.1

[6] http://cran.r-project.org/web/packages/mice/mice.pdf

Choosing K with the ICL criterion



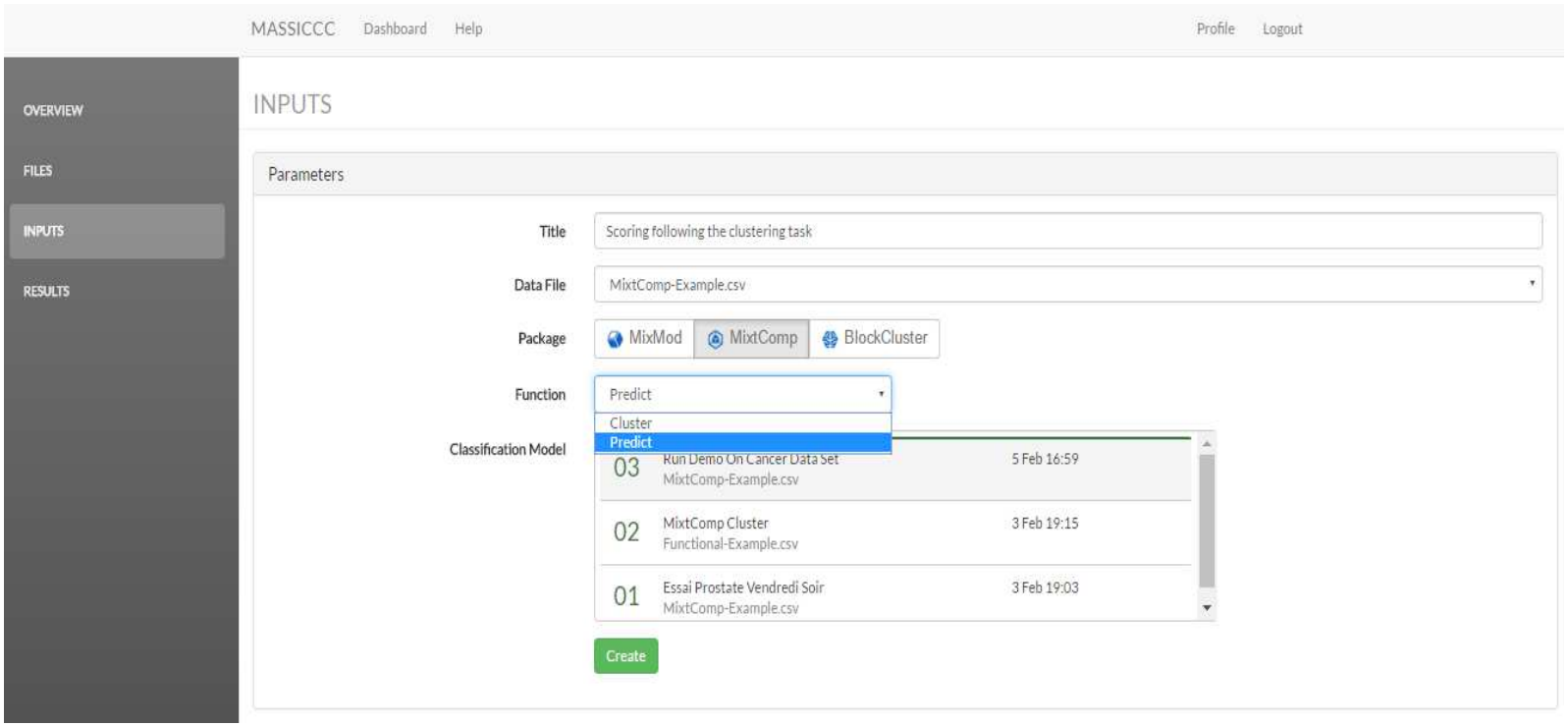
mice + MixtComp
$\hat{K} = 7$



full MixtComp
$\hat{K} = 2$

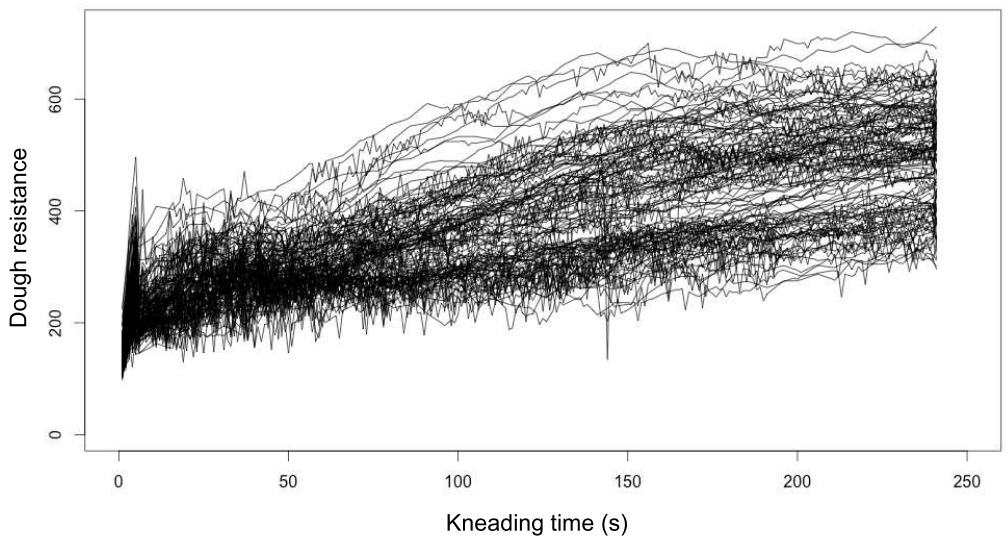
. . . may lose some cluster information when imputation before clustering

Scoring cancer data following the clustering task



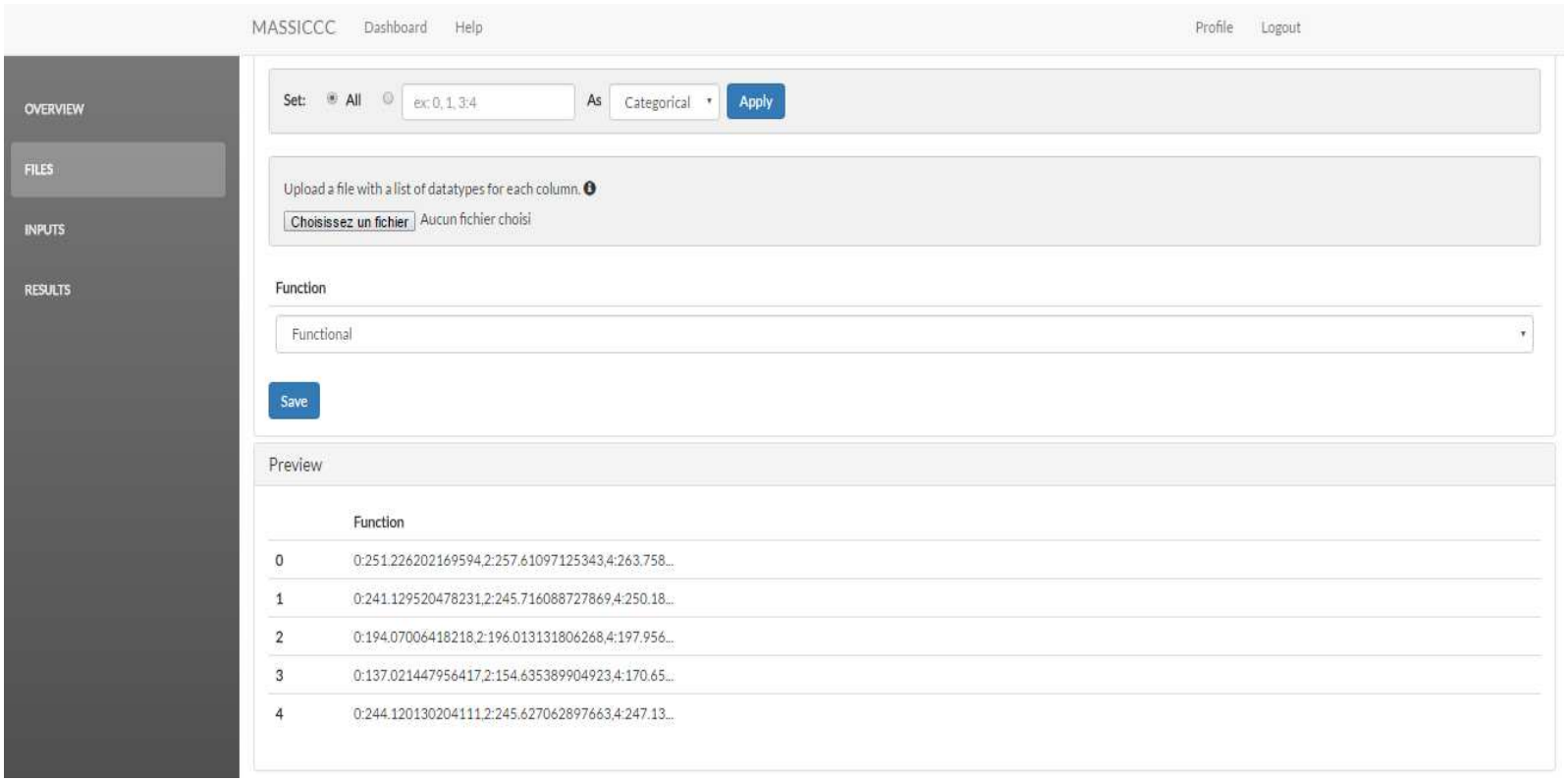
Curve "cookies" data set

The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process[7]. 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.

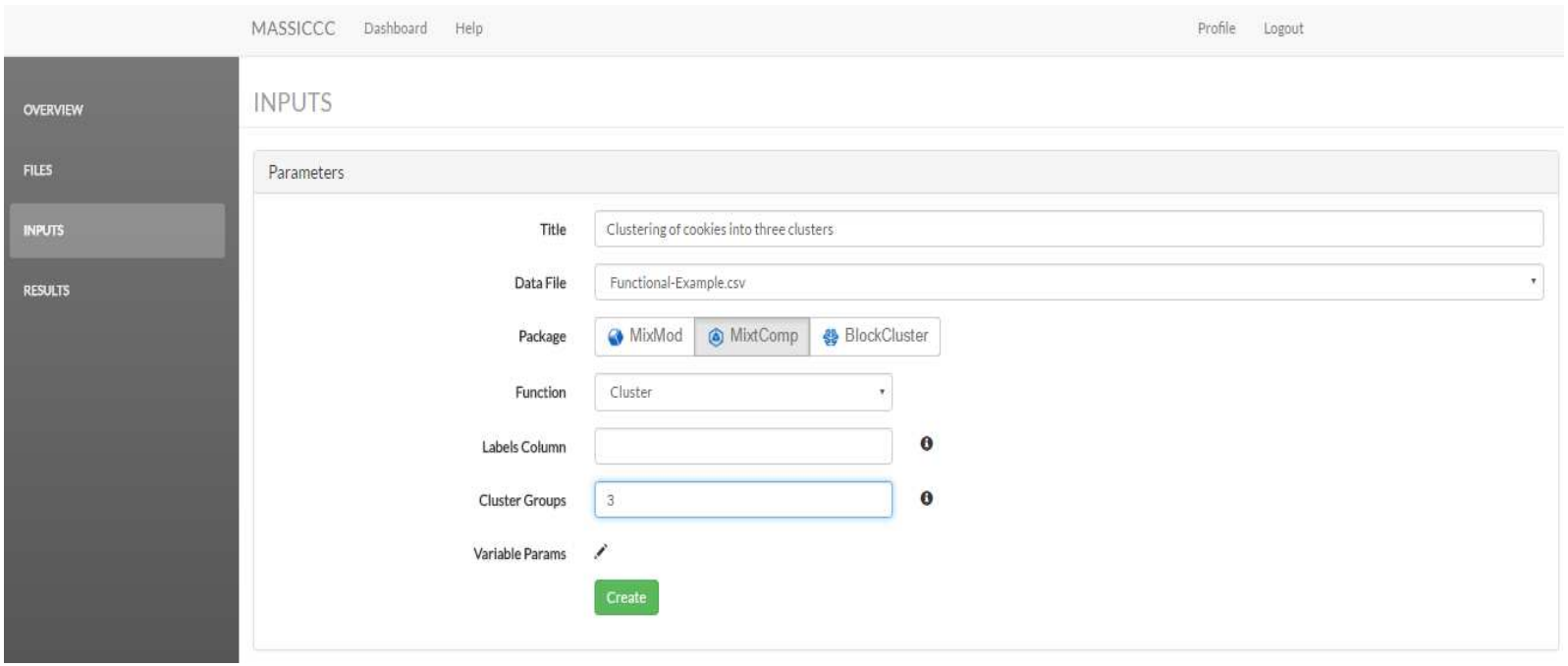


[7] Lévéder *et al*, 04

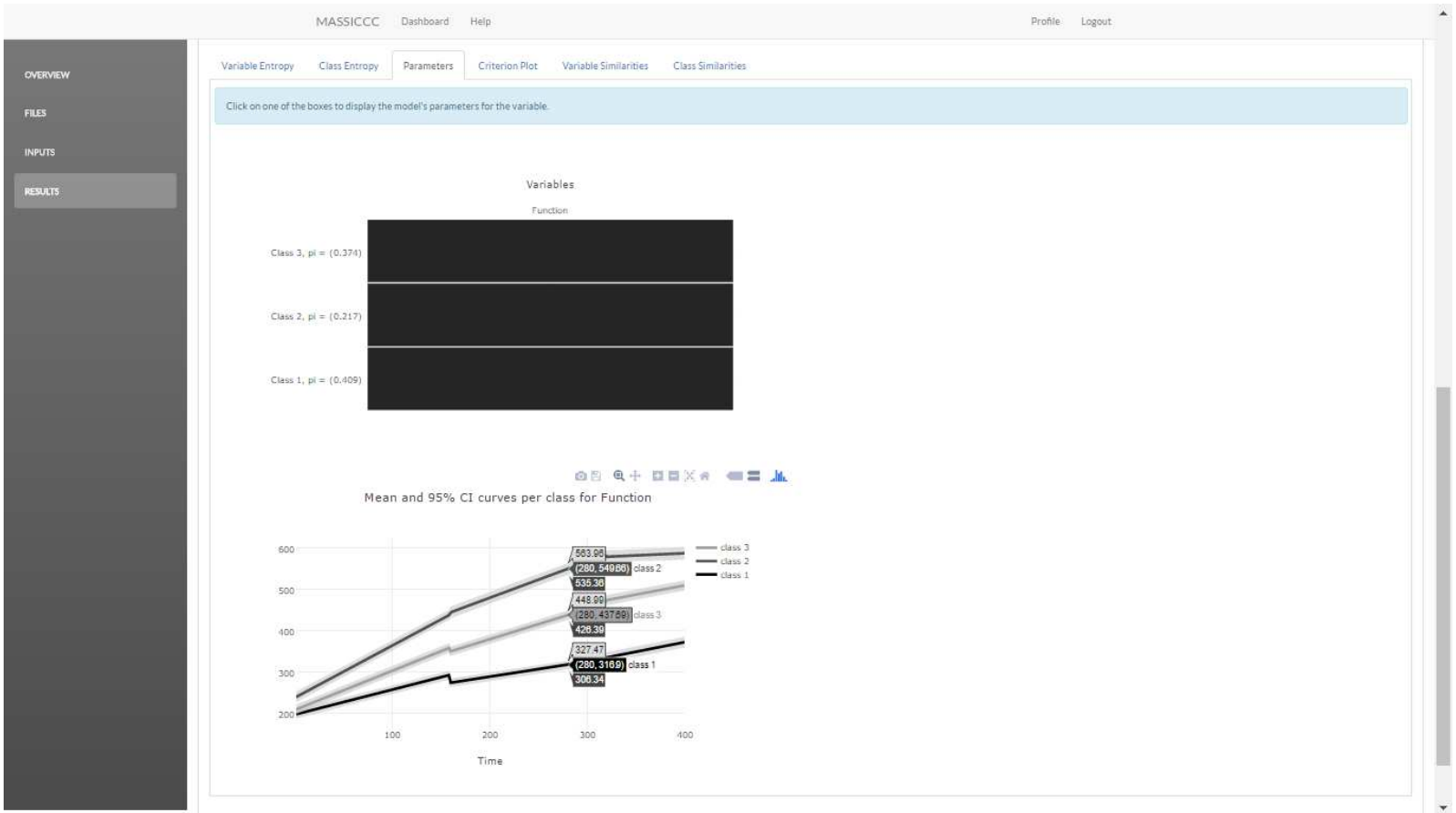
Upload curves data



Run a clustering task with three clusters

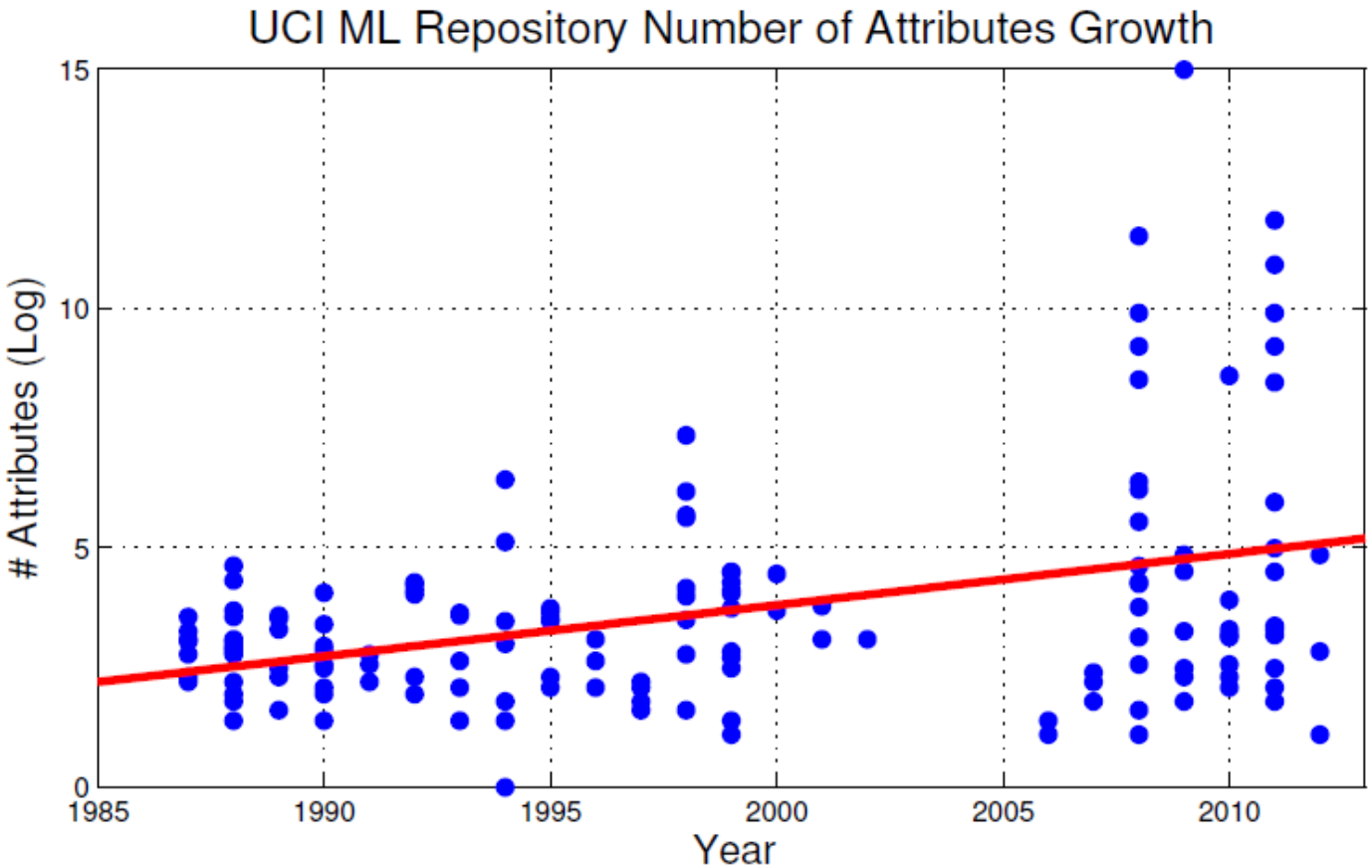


Overview of the three clusters of cookies



Outline

High-dimensional (HD) data[8]



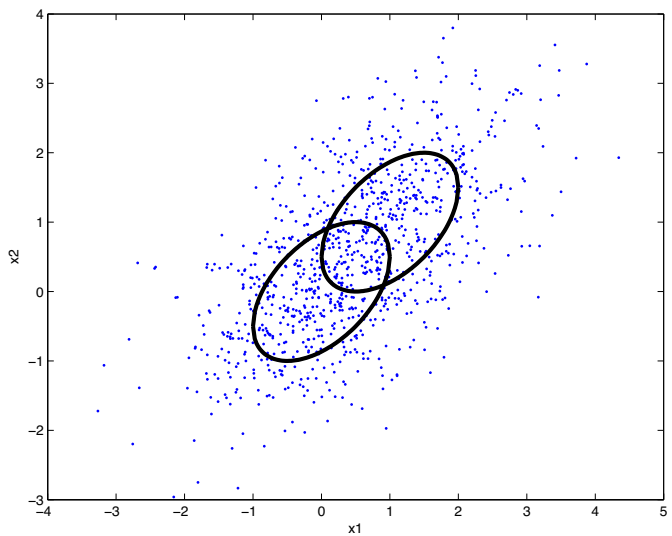
[8] S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

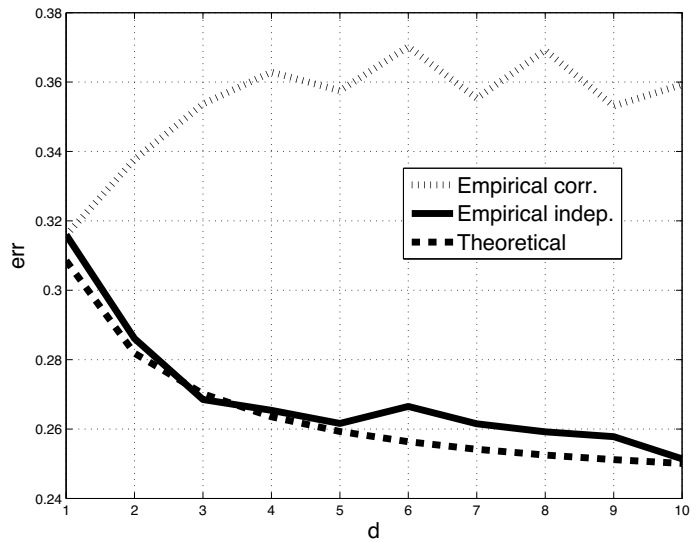
Bias/variance in HD: reduce variance, accept bias

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

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

- Each variable provides equal and own separation information
- Theoretical error decreases when d grows: $\mathrm{err}_{theo} = \Phi(-\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{\Sigma}^{-1}}/2)$
- 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)

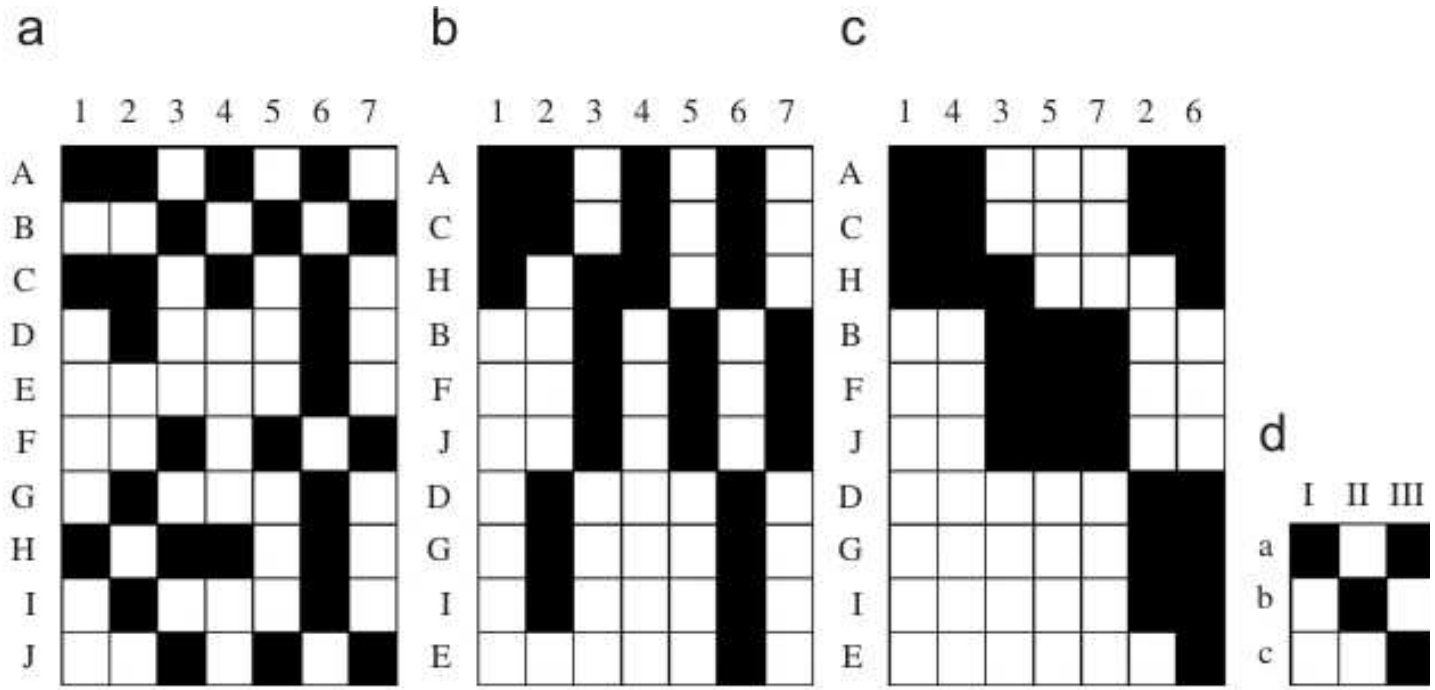
But which kind of parsimony?

- 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)

- $\mathbf{z} = (z_1, \ldots, z_n)$: partition of individuals in K custers of rows
- $\mathbf{w} = (w_1, \ldots, w_d)$: 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 ($\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$):

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

- $\boldsymbol{\pi} = (\pi_k)$: vectors of proba. π_k that a row belongs to the kth row cluster
- $\boldsymbol{\rho} = (\rho_k)$: vectors of proba. ρ_k that a row belongs to the lth column cluster
- Independence between all z_i and w_j
- Extension of the traditional mixture model-based clustering ($\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$):

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

Distribution for different kinds of data

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

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

[9] The 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 $\sum_k \pi_k \gamma_{kl} = \sum_l \rho_l \gamma_{kl} = 1$ 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	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + 2KL$

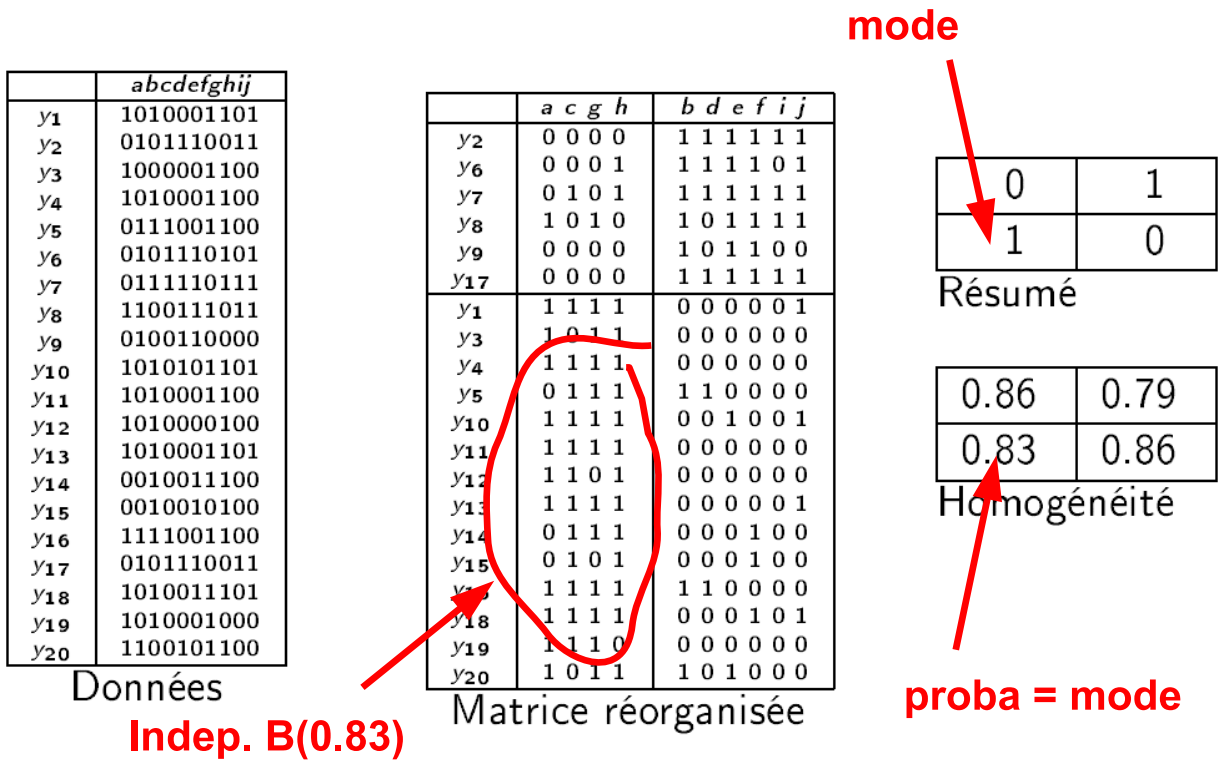
Very parsimonious so well suitable for the (ultra) HD setting

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

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

Binary illustration: easy interpretation

[Govaert, 2011]



	abcdefghij
y_1	1010001101
y_2	0101110011
y_3	1000001100
y_4	1010001100
y_5	0111001100
y_6	0101110101
y_7	0111110111
y_8	1100111011
y_9	0100110000
y_{10}	1010101101
y_{11}	1010001100
y_{12}	1010000100
y_{13}	1010001101
y_{14}	0010011100
y_{15}	0010010100
y_{16}	1111001100
y_{17}	0101110011
y_{18}	1010011101
y_{19}	1010001000
y_{20}	1100101100

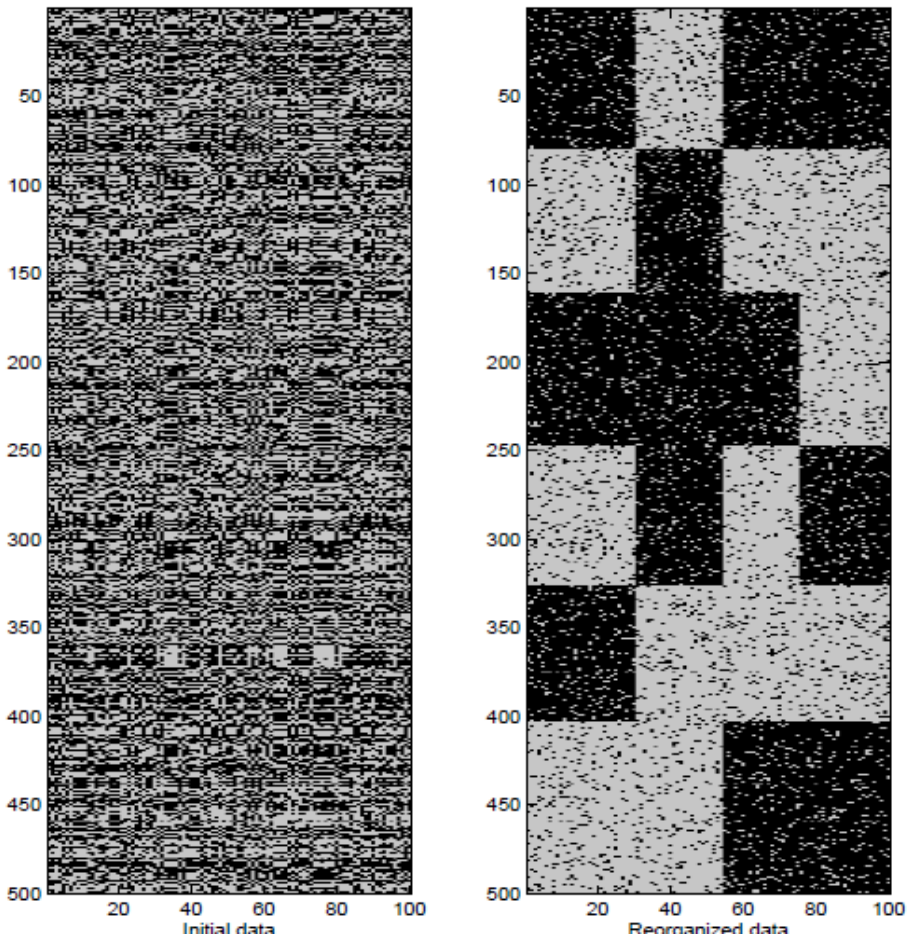
	a c g h	b d e f i j
y_2	0 0 0 0	1 1 1 1 1 1
y_6	0 0 0 1	1 1 1 1 0 1
y_7	0 1 0 1	1 1 1 1 1 1
y_8	1 0 1 0	1 0 1 1 1 1
y_9	0 0 0 0	1 0 1 1 0 0
y_{17}	0 0 0 0	1 1 1 1 1 1
y_1	1 1 1 1	0 0 0 0 0 1
y_3	1 0 1 1	0 0 0 0 0 0
y_4	1 1 1 1	0 0 0 0 0 0
y_5	0 1 1 1	1 1 0 0 0 0
y_{10}	1 1 1 1	0 0 1 0 0 1
y_{11}	1 1 1 1	0 0 0 0 0 0
y_{12}	1 1 0 1	0 0 0 0 0 0
y_{13}	1 1 1 1	0 0 0 0 0 1
y_{14}	0 1 1 1	0 0 0 1 0 0
y_{15}	0 1 0 1	0 0 0 1 0 0
[illegible]	1 1 1 1	1 1 0 0 0 0
y_{18}	1 1 1 1	0 0 0 1 0 1
y_{19}	1 1 1 0	0 0 0 0 0 0
y_{20}	1 0 1 1	1 0 1 0 0 0

0	1
1	0

0.86	0.79
0.83	0.86

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 $\boldsymbol{\theta}$, then deduce estimate of $(\boldsymbol{z}, \boldsymbol{w})$
- Observed log-likelihood: $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \ln \mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta})$
- MLE:

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

- Complete log-likelihood:

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

Be careful with asymptotics...

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

MLE estimation: EM algorithm

- E-step of EM (iteration q):

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

- 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 $\boldsymbol{x}_{ij}$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

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

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

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

  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 $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ or $\tilde{\ell}_c(\boldsymbol{\theta}; \boldsymbol{x}, \mathbf{e})$ are both intractable
- Idea: restriction on $\mathbf{e}$ to obtain tractability $\mathbf{e}_{ijkl} = t_{ik} s_{jl}$
- 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

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

MLE: variational EM (2/2)

Approximated E-step

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

- 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 $(\boldsymbol{z}^{(q)}, \boldsymbol{w}^{(q)}) \sim \mathrm{p}(\boldsymbol{z}, \boldsymbol{w}|\boldsymbol{x}; \boldsymbol{\theta}^{(q)})$ instead an expectation
- But it is still intractable, thus use a Gibbs algorithm to approx. this draw

Approximated S-step

Two easy draws

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

and

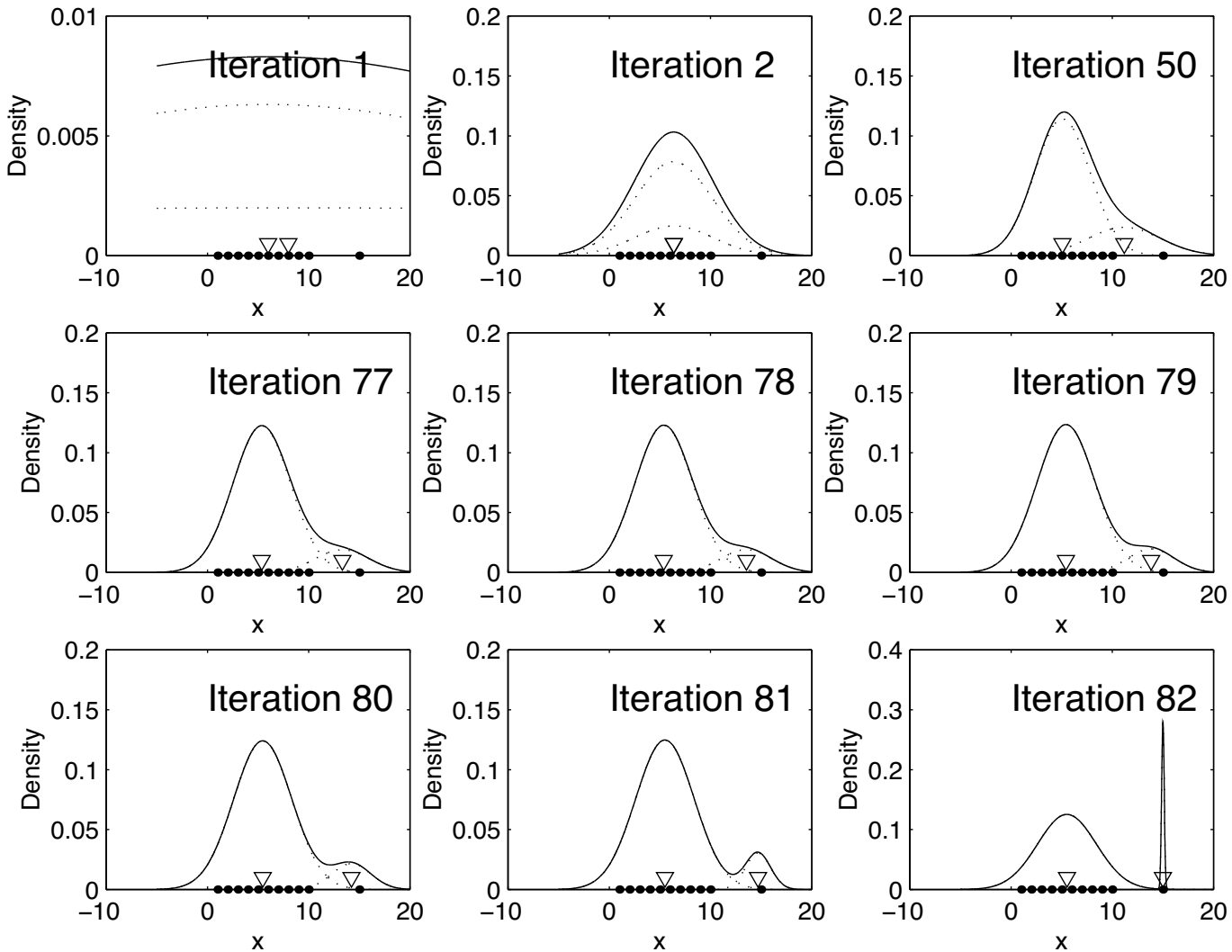
$$\boldsymbol{w}^{(q)} \sim \mathrm{p}(\boldsymbol{w}|\boldsymbol{z}^{(q)}, \boldsymbol{x}; \boldsymbol{\theta}^{(q)})$$

- 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 $\boldsymbol{\theta}$

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

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

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

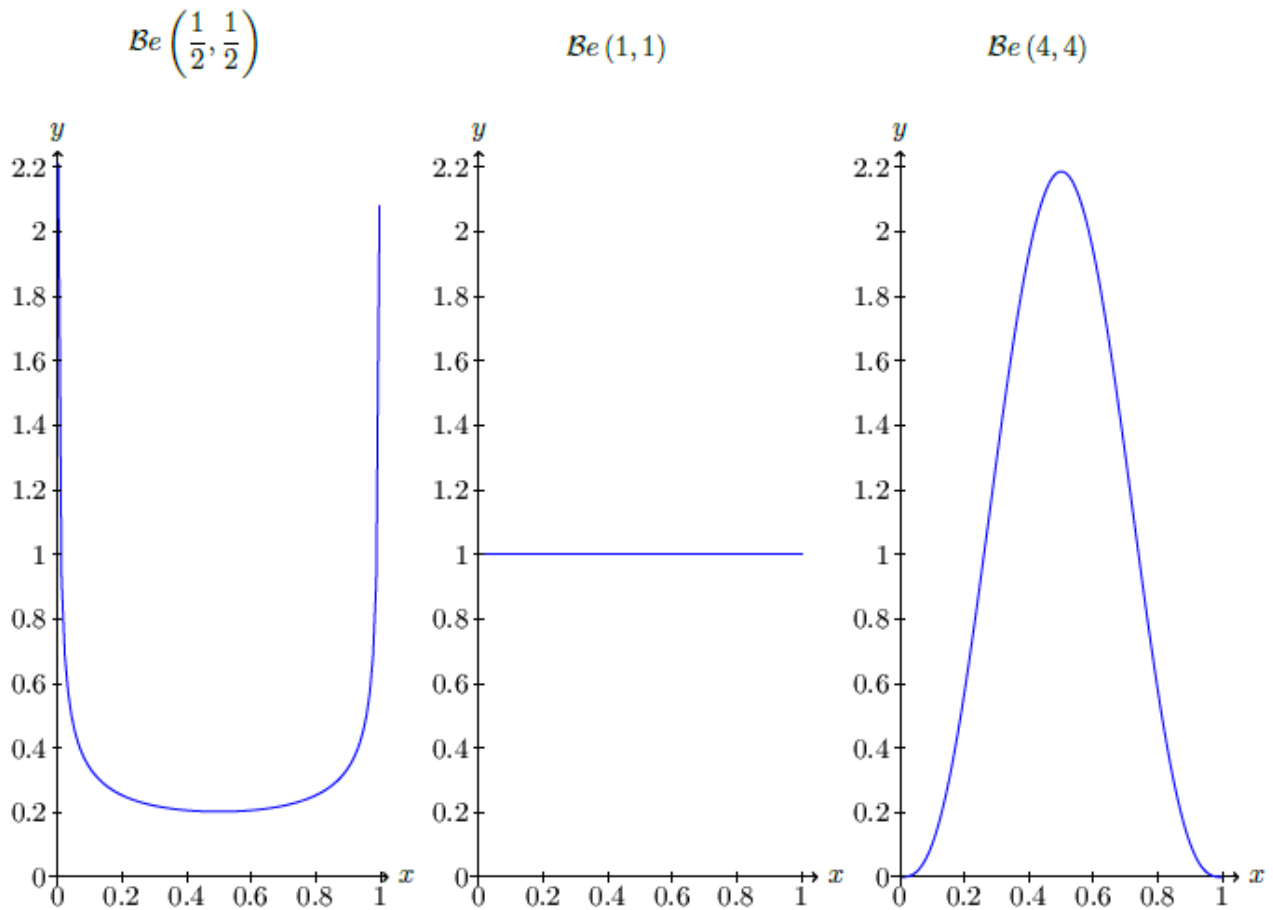
Bayesian estimation: limiting degeneracy

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

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

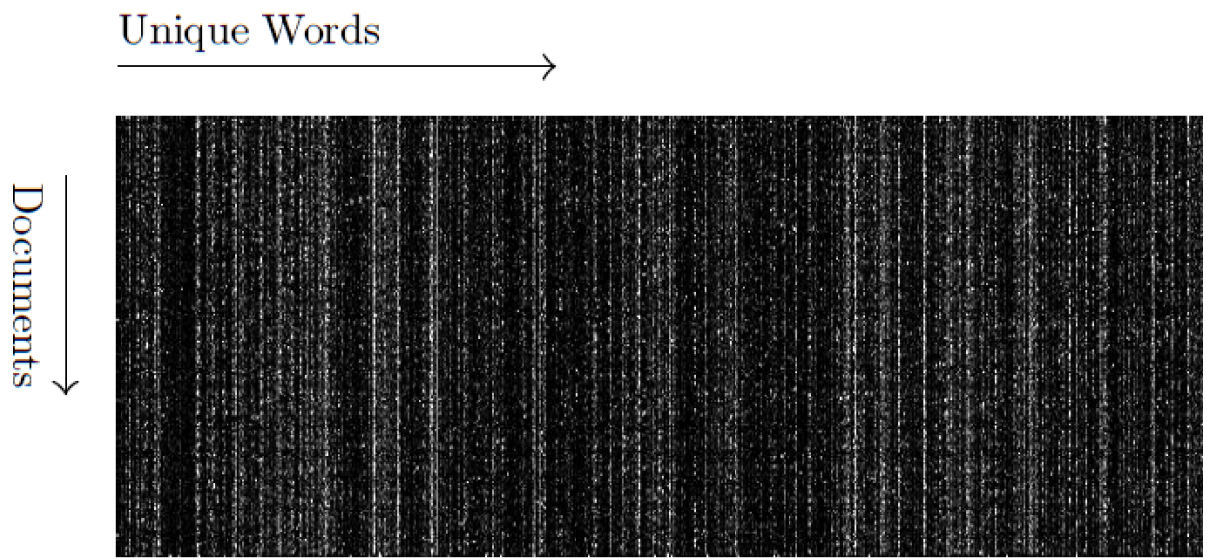
- The Dirichlet distribution is conjugate, thus easy calculus
- Control degeneracy frequency with the a value:
  - $a = 1$: uniform prior, so $\hat{\boldsymbol{\theta}}$ is strictly the MLE (no regularisation)
  - $a = 1/2$: Jeffreys prior, classical (no informative prior) but may favor degeneracy
 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

Bayesian estimation: prior overview



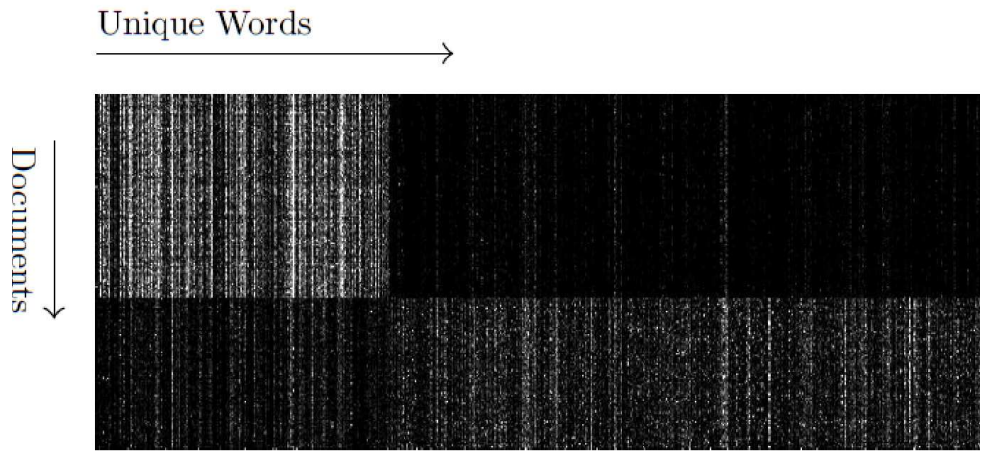
Document clustering (1/2)

- 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)

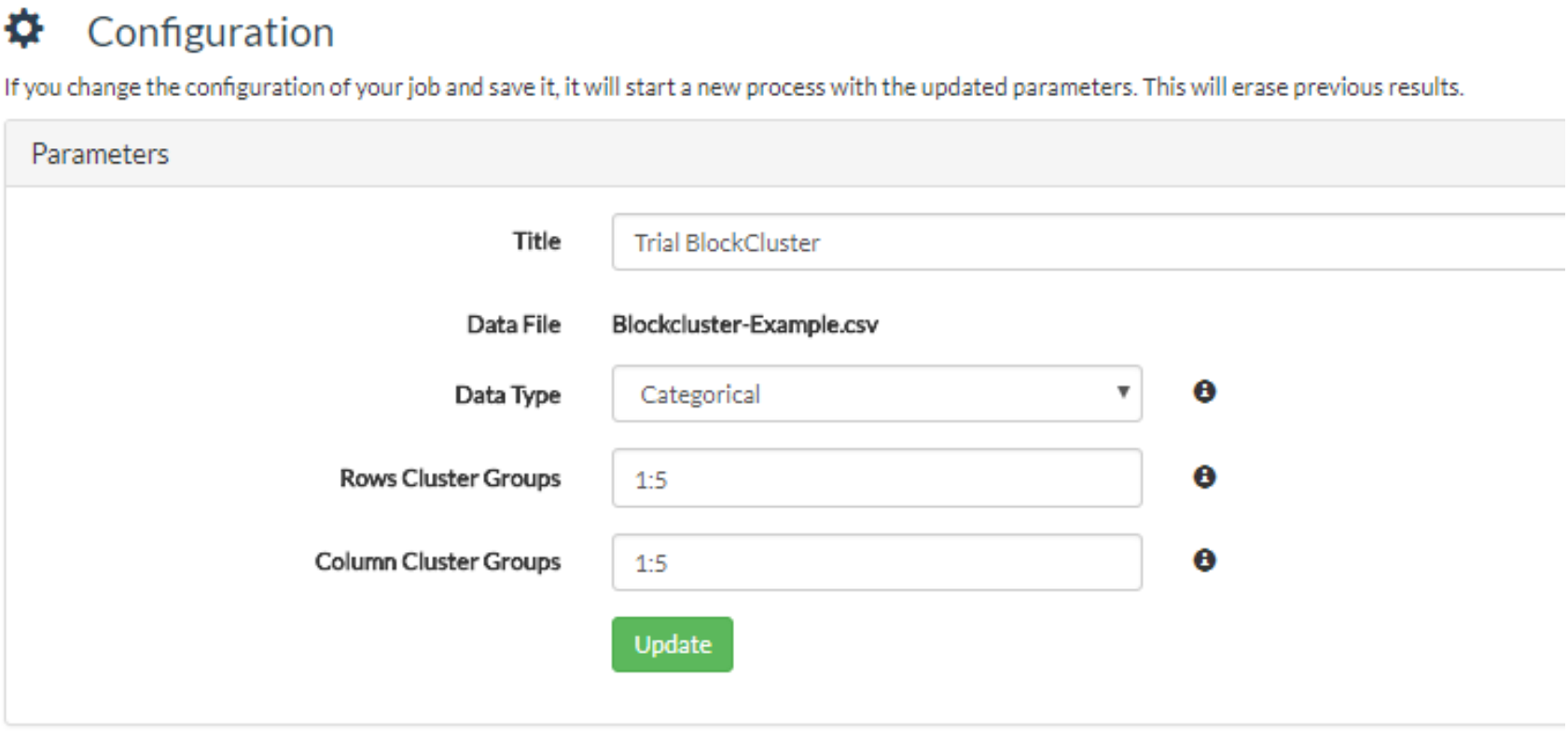
$$p(\hat{\mathbf{z}} \neq \mathbf{z}) \leq 2n \exp\left\{ -\frac{1}{8} d\Big[\underbrace{\min_{k \neq k'} |\tau_k - \tau_{k'}|}_{\text{overlap}} \Big] \right\} + K(1 - \min_k \pi_k)^n$$



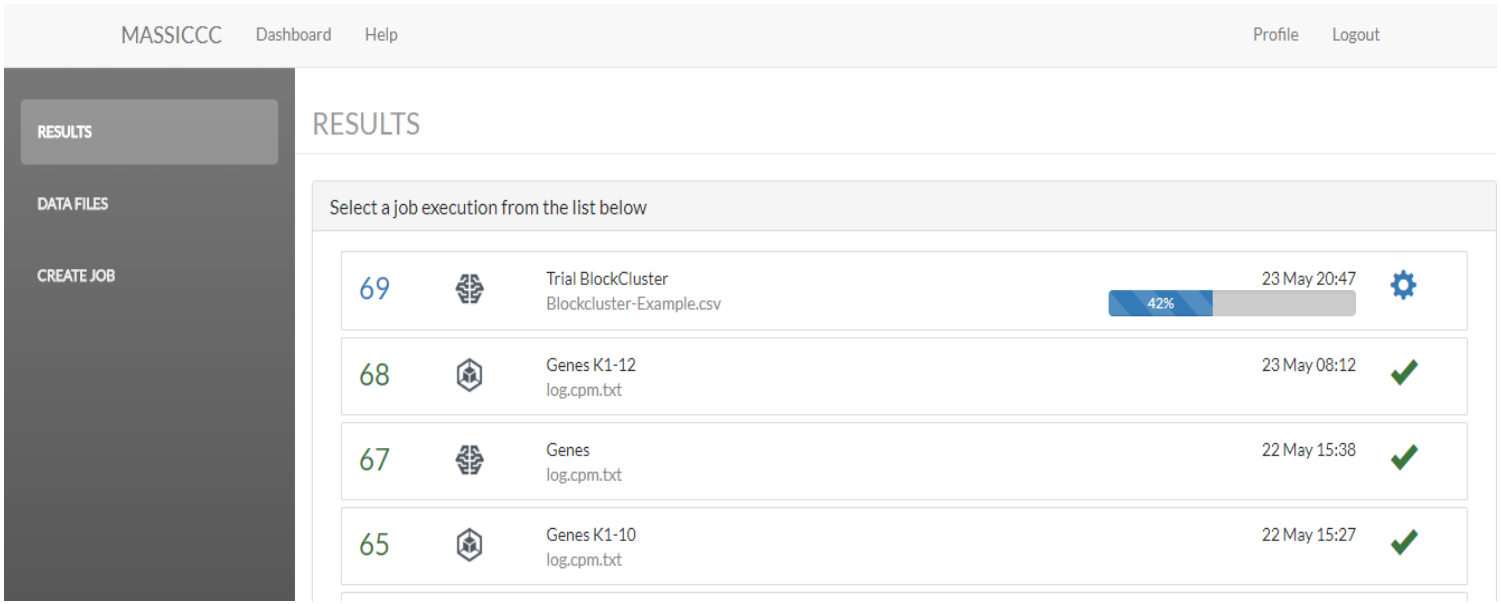
Results with 2×2 blocs

	Medline	Cranfield
Medline	1033	0
Cranfield	0	1398

Running BlockCluster



Running BlockCluster



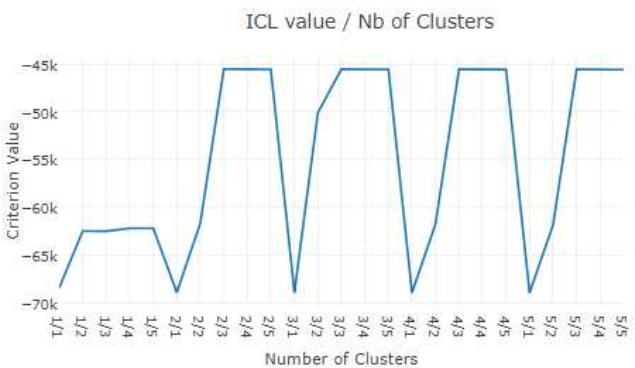
Running BlockCluster

Model	Criterion	Nb Clusters	Error
pik_rhol_multi	ICL (-45557.1)	[2,3]	No error
pik_rhol_multi	ICL (-45563.3)	[3,3]	No error
pik_rhol_multi	ICL (-45566.6)	[2,4]	No error
pik_rhol_multi	ICL (-45573.9)	[4,3]	No error
pik_rhol_multi	ICL (-45574.6)	[5,3]	No error
pik_rhol_multi	ICL (-45577.7)	[3,4]	No error
pik_rhol_multi	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.

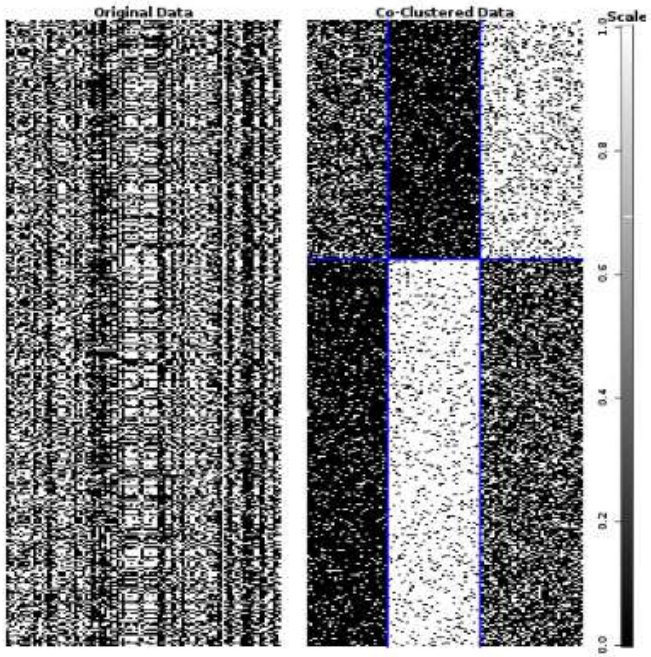


Running BlockCluster

Cluster Plot | Criterion Plot

Cluster Plot

This image presents the original data matrix, and the matrix obtained after performing co-clustering.



Outline

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

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