

MixtComp software: Model-based clustering/imputation with mixed data, missing data and uncertain data

<https://modal-research.lille.inria.fr/BigStat/>

Christophe Biernacki

(with Thibault Deregnacourt and Vincent Kubicki)

Tutorial in MissData Conference

June, 17th 2015



Take-home message

- **Imputation:** should take into account the final analysis purpose
- **Clustering:** no imputation is needed in the model-based context
- **Mixture models:** flexible enough for accurate multiple imputation

MixtComp software

Clustering/imputation for mixed data

Outline

1 Classifications(s): overview

2 Mixture model solution

3 Estimation

4 Clustering with MixtComp

5 Imputation with MixtComp

6 Conclusion

Today's data (1/2)

Today, it is easy to collect many features, so it favors

- data variety and/or mixed
- data missing
- data uncertainty (or interval data)

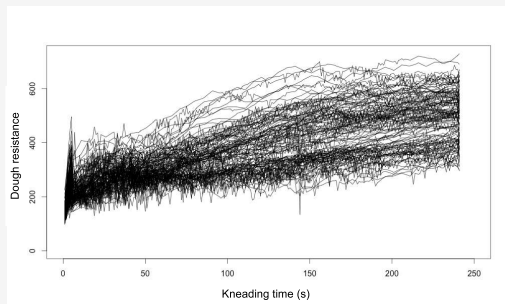
Mixed, missing, uncertain

Observed individuals $x^O \in \mathcal{X}$			
?	0.5	?	5
0.3	0.1	green	3
0.3	0.6	{red, green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

Today's data (2/2)

And also

- Ranking data
- Directional data
- Ordinal data
- Functional data
- Graphical data
- ...



Supervised classification (1/3)

■ **Data:** learning dataset $\mathcal{D} = (\mathbf{x}^O, \mathbf{z})$

- n individuals: $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) = (\mathbf{x}^O, \mathbf{x}^M)$ belonging to a space \mathcal{X}
- Observed individuals \mathbf{x}^O
- Missing individuals \mathbf{x}^M
- Partition in K groups G_1, \dots, G_K : $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$

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

■ **Aim:** estimation of an allocation rule r from \mathcal{D}

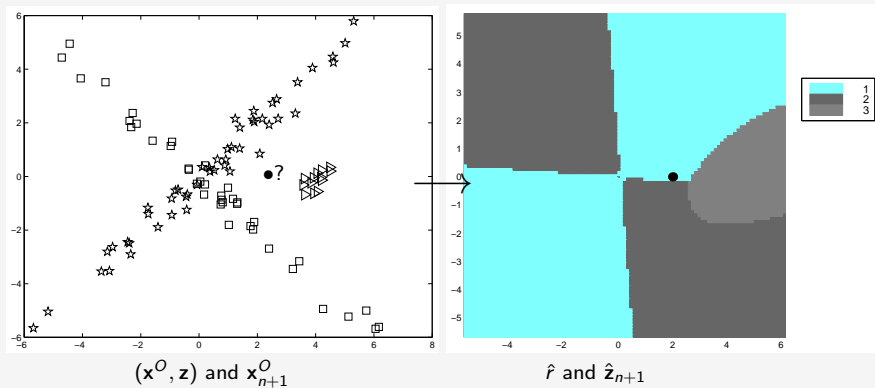
$$\begin{aligned} r : \quad \mathcal{X} &\longrightarrow \{1, \dots, K\} \\ \mathbf{x}_{n+1}^O &\longmapsto r(\mathbf{x}_{n+1}^O). \end{aligned}$$

Supervised classification (2/3)

Mixed, missing, uncertain

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

Supervised classification (3/3)



Semi-supervised classification (1/3)

- **Data:** learning dataset $\mathcal{D} = (\mathbf{x}^O, \mathbf{z}^O)$
 - n individuals: $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) = (\mathbf{x}^O, \mathbf{x}^M)$ belonging to a space \mathcal{X}
 - Observed individuals \mathbf{x}^O
 - Missing individuals \mathbf{x}^M
 - Partition: $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n) = (\mathbf{z}^O, \mathbf{z}^M)$
 - Observed partition \mathbf{z}^O
 - Missing partition \mathbf{z}^M
- **Aim:** estimation of an allocation rule r from \mathcal{D}

$$r : \begin{array}{ll} \mathcal{X} & \longrightarrow \{1, \dots, K\} \\ \mathbf{x}_{n+1}^O & \longmapsto r(\mathbf{x}_{n+1}^O). \end{array}$$

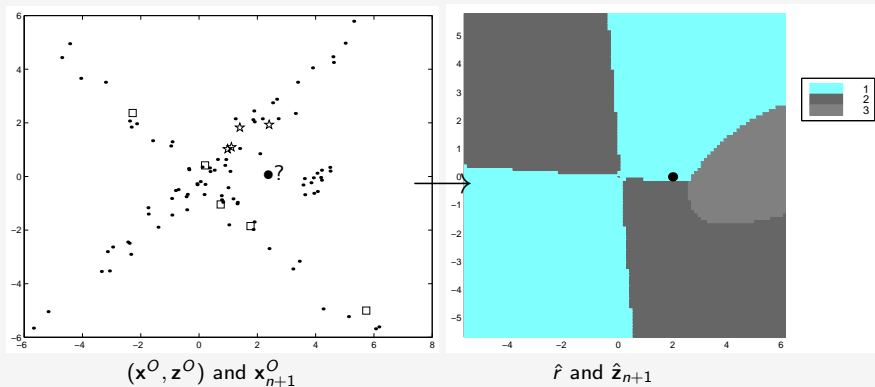
- **Idea:** \mathbf{x} is cheaper than \mathbf{z} so $\#\mathbf{z}^M \gg \#\mathbf{z}^O$

Semi-supervised classification (2/3)

Mixed, missing, uncertain

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

Semi-supervised classification (3/3)



Unsupervised classification (1/3)

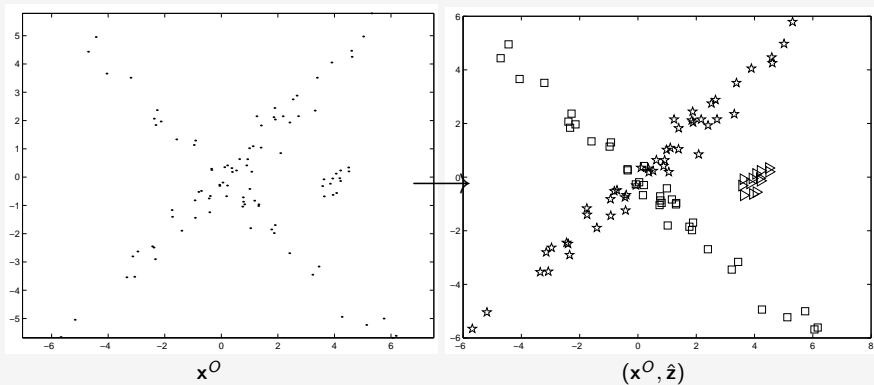
- **Data:** learning dataset $\mathcal{D} = \mathbf{x}^O$, so $\mathbf{z}^O = \emptyset$
- **Aim:** estimation of the partition \mathbf{z} and the number of groups K
- **Also known as:** clustering

Unsupervised classification (2/3)

Mixed, missing, uncertain

Individuals x^O				Partition z^O			\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					

Unsupervised classification (3/3)



Traditional solutions (1/3)

Two main frameworks

- **Generative models**

- Model $p(x, z)$
- Thus direct model for $p(x) = \sum_z p(x, z)$
- Easy to take into account some missing z and x

- **Predictive models**

- Model $p(z|x)$ or sometimes $\mathbf{1}_{\{p(z|x) > 1/2\}}$ or also ranking on $p(z|x)$
- Avoid assumptions on $p(x)$, thus avoids associated error model
- difficult to take into account some missing z and x

Traditional solutions (2/3)

No mixed, missing or uncertain data:

- **Supervised classification**¹
 - **Generative models:** linear/quadratic discriminant analysis
 - **Predictive models:** logistic regression, support vector machines (SVM), k nearest neighbourhood, classification trees. . .
- **Semi-supervised classification**²
 - **Generative models:** mixture models
 - **Predictive models:** low density separation (transductive SVM), graph-based methods. . .
- **Unsupervised classification**³
 - **Generative models:** k -means like criteria, hierarchical clustering, mixture models
 - **Predictive models:** -

¹Govaert *et al.*, Data Analysis, Chap.6, 2009

²Chapelle *et al.*, Semi-supervised learning, 2006

³Govaert *et al.*, Data Analysis, Chap.7-9, 2009

Traditional solutions (3/3)

But more complex with mixed, missing or uncertain data. . .

- **Missing/uncertain data**: multiple imputation is possible but it should ideally take into account the classification purpose at hand
- **Mixed data**: some heuristic methods with recoding

How to marry the classification aim with mixed, missing or uncertain data?

Outline

1 Classifications(s): overview

2 Mixture model solution

3 Estimation

4 Clustering with MixtComp

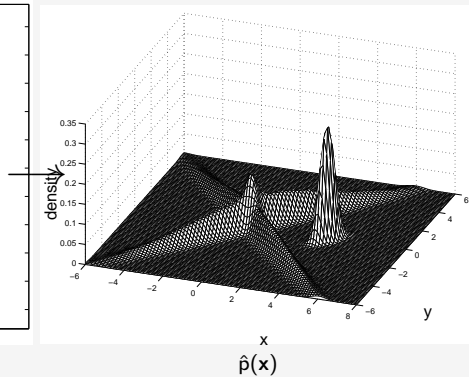
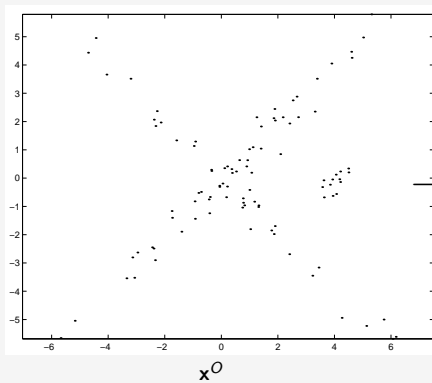
5 Imputation with MixtComp

6 Conclusion

Density estimation (1/2)

- **Data:** learning dataset $\mathcal{D} = \mathbf{x}^O$, so $\mathbf{z}^O = \emptyset$
- **Aim:** estimation of the distribution $p(\mathbf{x})$
- **Extension easy to:** $\mathcal{D} = (\mathbf{x}^O, \mathbf{z}^O)$ with $\mathbf{z}^O \neq \emptyset$
- **Useful for:** data imputation and multi-purpose classification!

Density estimation (2/2)



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

- Rigorous definition of a group:

$$\mathbf{x}_1 \in G_k \Leftrightarrow \mathbf{x}_1 \text{ is a realization of } \mathbf{X}_1 \sim p_k(\mathbf{x}_1)$$

- Mixture formulation:

$$\begin{aligned} \mathbf{X}_1 |_{Z_{1k}=1} &\sim p_k(\mathbf{x}_1) \\ \mathbf{Z}_1 &\sim \text{Mult}_K(1, \underbrace{\pi_1, \dots, \pi_K}_{\pi}) \end{aligned}$$

- Joint and marginal (or mixture) distributions:

$$\begin{aligned} (\mathbf{X}_1, \mathbf{Z}_1) &\sim \prod_{k=1}^K [\pi_k p_k(\mathbf{x}_1)]^{z_{1k}} \\ \mathbf{X}_1 &\sim p(\mathbf{x}_1) = \sum_{k=1}^K \pi_k p_k(\mathbf{x}_1) \end{aligned}$$

- Maximum a posteriori (MAP): with $t_k(\mathbf{x}_1^O) = p(Z_{1k} = 1 | \mathbf{x}_1^O) = \frac{\pi_k p_k(\mathbf{x}_1^O)}{p(\mathbf{x}_1^O)}$

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

The mixture model answer for imputation

Straightforward also, for instance by the mode

$$\hat{x}^M = \arg \max_{x^M} p(x^M | x^O)$$

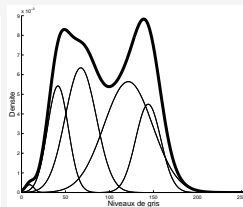
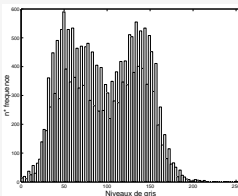
Other possibilities, depending on the data type: mean, *etc.*

Distribution $p(x^M | x^O)$

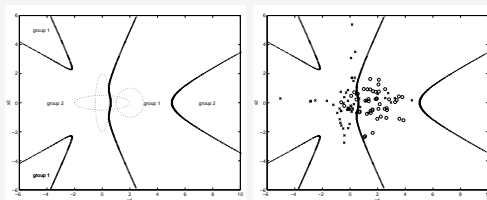
It allows also to perform a specific multiple imputation!

The mixture model answer in density estimation

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



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



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

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

Mixed data: conditional independence everywhere

The aim is to combine continuous, categorical and integer data

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

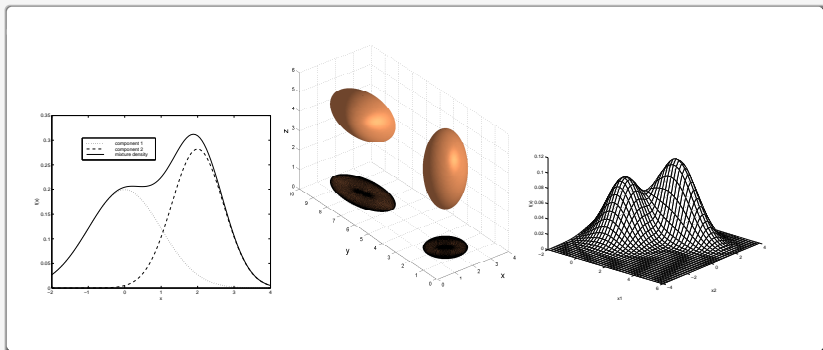
The proposed solution is to mixed all types by **inter conditional independence**

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

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

Continuous: Gaussian mixture model

$$p(\cdot; \alpha_k^{cont}) = N_d(\mu_k, \underbrace{\Sigma_k}_{\text{diagonal}})$$



Categorical: latent class model

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

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

- **Intra conditional independence:**

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

and

$$\alpha_k^{jh} = p(X_i^{jh} = 1 | Z_{ik} = 1)$$

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

Integer: Poisson mixture model

- **integer variables:** d variables $\mathbf{x}_i^j \in \mathbb{N}$
- **Intra conditional independence:**

$$p(\mathbf{x}_i^{int}; \boldsymbol{\alpha}_k^{int}) = \prod_{j=1}^d \frac{(\alpha_k^j)^{x_i^j}}{\alpha_k^j!} e^{-\alpha_k^j}$$

Outline

1 Classifications(s): overview

2 Mixture model solution

3 Estimation

4 Clustering with MixtComp

5 Imputation with MixtComp

6 Conclusion

Sampling assumptions

- True distribution:

$$\mathcal{D} \sim p(\mathcal{D})$$

- Model distribution:

$$(\mathbf{x}_i, \mathbf{z}_i) \stackrel{i.i.d.}{\sim} p(\mathbf{x}_1, \mathbf{z}_1; \boldsymbol{\theta})$$

- Gap between both, but **flexibility**:

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \Theta} \text{KL}(p, p_{\boldsymbol{\theta}})$$

where

$$\text{KL}(p, p_{\boldsymbol{\theta}}) = E_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ln p(\mathcal{D}'; \boldsymbol{\theta})]$$

Observed-data log-likelihood estimation of θ

- **Principle:** MLE

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \ell(\theta; \mathcal{D})$$

with observed log-likelihood

$$\ell(\theta; \mathcal{D}) = \ln p(\mathcal{D}; \theta) = \ln \int_{\mathbf{x}^M} \sum_{\mathbf{z}^M} p(\mathbf{x}, \mathbf{z}; \theta) d\mathbf{x}^M$$

- **Consistency:** we have

$$\hat{\theta} \xrightarrow{a.s.} \theta^*$$

- **Algorithm:** SEM

SEM algorithm

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

Properties

- simplicity because of conditional independence
- 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

SE algorithm

A SE algorithm estimates then $(\mathbf{x}^M, \mathbf{z}^M)$

- Iteration nb q :
 - **E-step**: compute conditional probabilities $p(\mathbf{x}^M, \mathbf{z}^M | \mathcal{D}; \hat{\theta})$
 - **S-step**: draw $(\mathbf{x}^{M(q)}, \mathbf{z}^{M(q)})$ from $p(\mathbf{x}^M, \mathbf{z}^M | \mathcal{D}; \hat{\theta})$
- Stopping rule: iteration number

Properties

- simplicity because of conditional independence
- the mean/mode of the sequence $(\mathbf{x}^{M(q)}, \mathbf{z}^{M(q)})$ estimates $(\mathbf{x}^M, \mathbf{z}^M)$
- confidence intervals are also derived

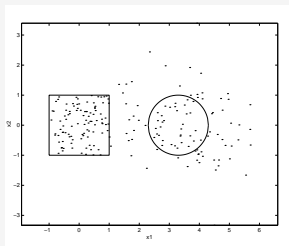
Estimating K

- Density estimation purpose:

$$\text{BIC} = \ln p(\mathbf{x}^O, \mathbf{z}^O; \hat{\boldsymbol{\theta}}) - \frac{\text{nb param.}}{2} \ln(n)$$

- Clustering purpose:

$$\text{ICL} = \ln p(\mathbf{x}^O, \mathbf{z}^O, \hat{\mathbf{z}}^M; \hat{\boldsymbol{\theta}}) - \frac{\text{nb param.}}{2} \ln(n)$$



\hat{K}	1	2	3	4	5
BIC	.	60	.	32	8
ICL	.	100	.	.	.

What about the process that causes missing data?

Biometrika (1976), **63**, 3, pp. 581–92

Printed in Great Britain

581

Inference and missing data

BY DONALD B. RUBIN

Educational Testing Service, Princeton, New Jersey

SUMMARY

When making sampling distribution inferences about the parameter of the data, θ , it is appropriate to ignore the process that causes missing data if the missing data are ‘missing at random’ and the observed data are ‘observed at random’, but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ , it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is ‘distinct’ from θ . These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.

Outline

1 Classifications(s): overview

2 Mixture model solution

3 Estimation

4 Clustering with MixtComp

5 Imputation with MixtComp

6 Conclusion

Prostate cancer data⁴ (1/2)

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

⁴Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Prostate cancer data (2/2)

<i>Covariate</i>	<i>Abbreviation</i>	<i>Number of Levels</i> (if categorical)
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histologic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2

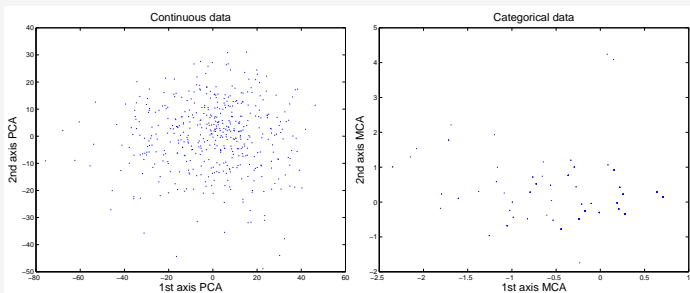
Aim

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

Questions

- How many clusters?
- Which partition?

Visually not so easy...



Create an account in MixtComp

<https://modal-research.lille.inria.fr/BigStat/>

BigStat MixtComp HDPenReg MixAll BlockCluster Dev Login Register

Log in

Username
biemacki

Password

Log In

[Forgot Password?](#)

See documentation at <https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp>

Variable descriptor file: descriptor.csv

The screenshot shows the OpenOffice Calc interface with the file 'prostate_descriptor.csv' open. The spreadsheet contains a table with 13 columns (A-M) and 3 rows. The first row (A1) is the header 'z_class'. The second row (A2) lists the variable names: 'Age', 'Wt', 'PF', 'HX', 'SBP', 'DBP', 'EKG', 'HG', 'SZ', 'SG', 'AP', and 'BM'. The third row (A3) lists the corresponding distributions: 'LatentClass', 'Gaussian_sjk', 'Gaussian_sjk', 'Categorical_pjk', 'Categorical_pjk', 'Gaussian_sjk', 'Gaussian_sjk', 'Categorical_pjk', 'Gaussian_sjk', 'Gaussian_sjk', 'Gaussian_sjk', 'Gaussian_sjk', and 'Categorical_pjk'.

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

Syntax/allowed missing data

allowed missing value types for each model

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

Data file: data.csv

data.csv - Op

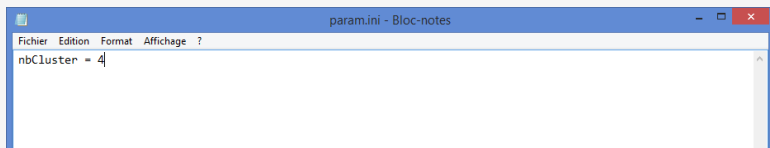
Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

Arial 10 G I S

A1 z_class

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	z_class	Age	Wt	PF	HX	SBP	DBP	EKG	HG	SZ	SG	AP	BM	
2	?	75	76	1	1	15	9	5	138	1.4142	8	1.0986	1	
3	?	54	116	1	1	13	7	4	146	6.4807	?	1.9459	1	
4	?	69	102	1	2	14	8	5	134	1.7321	9	1.0986	1	
5	?	75	94	2	2	14	7	2	176	2	82	1.972	1	
6	?	67	99	1	1	17	10	1	134	5.831	8	1.6094	1	
7	?	71	98	1	1	19	10	1	151	3.1623	11	1.7918	1	
8	?	75	100	1	1	14	10	2	130	3.6056	9	2.0794	1	
9	?	73	114	1	2	17	11	5	126	1.7321	9	1.7918	1	
10	?	60	110	1	1	12	8	1	146	2	10	1.9459	1	
11	?	78	107	1	2	13	8	6	130	4.5826	6	1.3863	1	
12	?	77	89	1	1	15	8	1	156	1.7321	8	1.7918	1	
13	?	74	105	1	2	18	14	1	136	2.4495	8	1.3863	1	
14	?	74	107	1	1	14	9	6	144	2.4495	9	1.0986	1	
15	?	55	112	1	2	16	9	5	139	2	9	2.3026	1	
16	?	73	88	1	1	19	10	5	120	3.873	10	1.7918	1	
17	?	87	81	2	2	17	12	3	134	1.7321	9	1.3863	1	
18	?	64	90	1	1	14	8	1	162	2.4495	9	1.9459	1	
19	?	79	104	1	1	13	8	2	150	2.2361	8	1.6094	1	
20	?	62	90	1	2	13	8	2	144	1.4142	9	1.9459	1	
21	?	74	107	1	1	14	9	6	144	2.4495	9	1.0986	1	

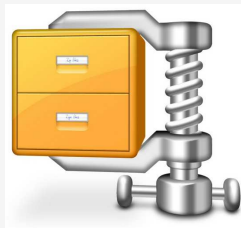
Number of clusters file: param.ini



```
param.ini - Bloc-notes
Fichier Edition Format Affichage ?
nbCluster = 4
```

Input file: *.zip

descriptor.csv
+
data.csv
+
param.ini
=
NameYouWant.zip



Learn!

BigStat MixtComp ▾ HDPenReg ▾ Mixture ▾ BlockCluster ▾ Dev ▾ [Home](#) [Logout](#)

Learn

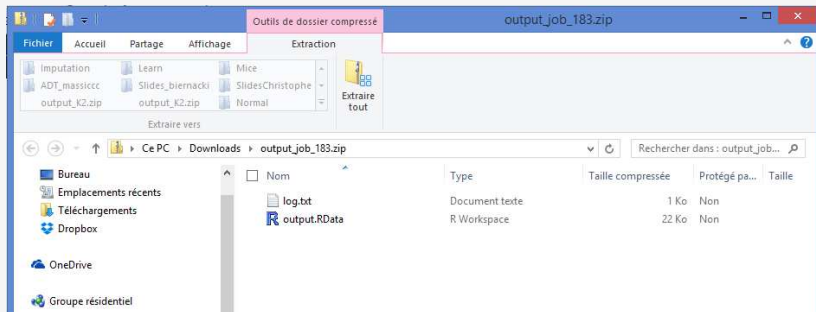
Upload PData
20.1 Mo 17 juin 2015 11:14

Tutorial NameYouWant.zip

First 1 2 Last

Title	Status		Creation	Begin	End	Downloads
Tutorial	43%	✖	16 juin 2015 11:25:55	16 juin 2015 11:25:55		<input type="button" value="Input"/>
essaiK1	Completed	🗑️	16 juin 2015 11:14:34	16 juin 2015 11:14:34	16 juin 2015 11:14:36	<input type="button" value="Input"/> <input type="button" value="Output"/> <input type="button" value="Log"/>

Output zip file



Output R format

```

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

```

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

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

Two strategies in competition

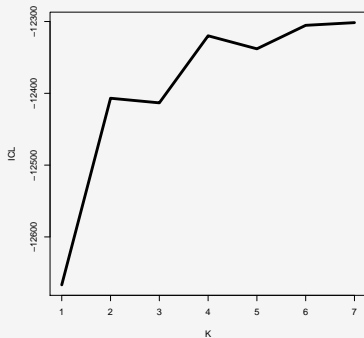
- **Strategy “mice⁵ + MixtComp”**: MixtComp on the dataset completed by mice

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

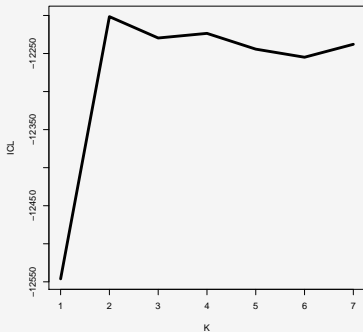
- **Strategy “full MixtComp”**: MixtComp on the observed (no completed) dataset

⁵<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

Partition quality with $K = 2$

Strategy	mice + MixtComp	full MixtComp
% misclassified	12.8	8.1

To be compared also to missing data removal:

- 475 patients with non-missing data
- MixtComp for clustering
- possibility to consider continuous, categorical or mixed data

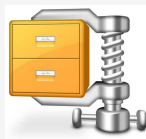
Strategy	continuous only	categorical only	mixed cont/cat
% misclassified	9.46	47.16	8.63

- risk of information lost when removing missing data lines/columns
- avoid to complete missing data (**imputation depends on the purpose**)

And for supervised classification?

Use now the [predict](#) functionality of MixtComp

descriptor.csv
+
data.csv
+
[output.RData](#)
(from previous learn. . .)
=
NameYouWant.zip



Then same output format as the [learn](#) functionality of MixtComp

Outline

1 Classifications(s): overview

2 Mixture model solution

3 Estimation

4 Clustering with MixtComp

5 Imputation with MixtComp

6 Conclusion

Cancer dataset with more missing data

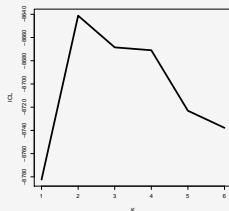
Add artificially $\approx 30\%$ missing data with a MCAR design

Then compare two strategies of imputation:

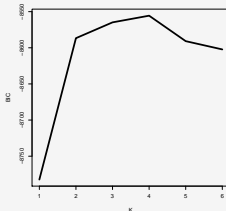
- Strategy “mice”: dataset completed by mice

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

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



ICL
 $\hat{K} = 2$



BIC
 $\hat{K} = 4$

Output multiple imputation by MixtComp

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

[[2]]
[1] 70.62032

[[3]]
[1] 58.24255

[[4]]
[1] 83.86463
```

cont.

```
> res$variable$data$SEKG$completed[[5]]
[1] 1
> res$variable$data$SEKG$stat[[1]]
[[1]]
[1] 5

[[2]]
[1] 1

[[3]]
[1] 0.41

[[4]]
[1] 5

[[5]]
[1] 0.29

[[6]]
[1] 6

[[7]]
[1] 0.15

[[8]]
[1] 2

[[9]]
[1] 0.07

[[10]]
[1] 3

[[11]]
[1] 0.05
```

cat.

Imputation accuracy

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

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

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

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

Outline

- 1 Classifications(s): overview
- 2 Mixture model solution
- 3 Estimation
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp
- 6 Conclusion**

Conclusion

- **Clustering**: work directly on observed (not imputed) data
- **Imputation**: possible since flexibility of mixture models for density estimation
- **MixtComp**: clustering and/or imputation for mixed data
 - **Now**: continuous, categorical, integer
 - **Next**: ordinal, ranks, functional, directional