

Model-based clustering with mixed/missing data using the new software MixtComp

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

Christophe Biernacki

(with Thibault Deregnacourt and Vincent Kubicki)

CMStatistics 2015 (ERCIM 2015)
London (UK), 12-14 December 2015



Outline

- 1** The problem
- 2 Conditional independent clustering
- 3 Estimation
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp
- 6 Conclusion

Clustering of complex data

- **Data:** 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
- **Aim:** estimation of the partition \mathbf{z} and the number of clusters K
 Partition in K clusters G_1, \dots, G_K : $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})'$

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

Mixed, missing, uncertain

Individuals \mathbf{x}^O				Partition \mathbf{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			

Model-based clustering

- Cluster k is modelled by a **parametric distribution**: $\mathbf{X}_i|Z_{ik}=1 \stackrel{i.i.d.}{\sim} p(\cdot; \boldsymbol{\alpha}_k)$
- Cluster k has probability π_k with $\sum_{k=1}^K \pi_k = 1$: $\mathbf{Z}_i \stackrel{i.i.d.}{\sim} \text{Mult}_K(1, \pi_1, \dots, \pi_K)$
- Missing data \mathbf{x} are obtained by a **missing completely at random process (MCAR)**¹
- Observed mixture pdf**: with parameter $\boldsymbol{\theta} = (\pi_1, \dots, \pi_K, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K)$, it is written

$$p(\mathbf{x}_i^O; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k p(\mathbf{x}_i^O; \boldsymbol{\alpha}_k) = \sum_{k=1}^K \pi_k \int_{\mathbf{x}_i^M} p(\mathbf{x}_i^O, \mathbf{x}_i^M; \boldsymbol{\alpha}_k) d\mathbf{x}_i^M$$

- Maximum a posteriori (MAP)**: with $t_k(\mathbf{x}_i^O; \boldsymbol{\theta}) = p(Z_{ik} = 1 | \mathbf{x}_i^O; \boldsymbol{\theta}) = \frac{\pi_k p(\mathbf{x}_i^O; \boldsymbol{\alpha}_k)}{p(\mathbf{x}_i^O; \boldsymbol{\theta})}$

$$\hat{\mathbf{z}}_i = \arg \max_{k=\{1, \dots, K\}} t_k(\mathbf{x}_i^O; \boldsymbol{\theta})$$

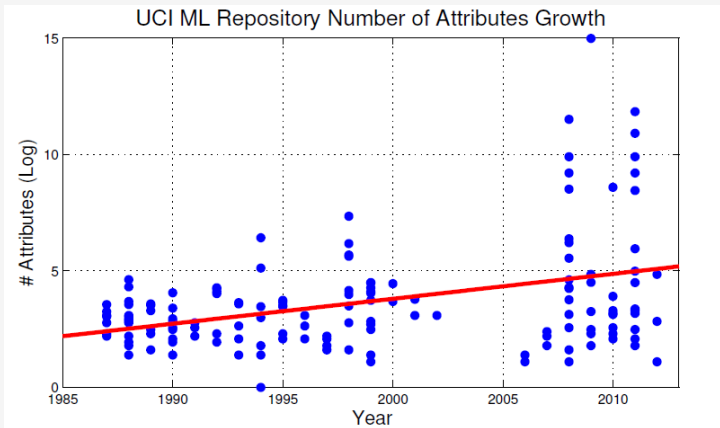
Seems to be suitable for missing/uncertain data but **which $p(\cdot; \boldsymbol{\alpha}_k)$** for mixed data?

¹Could be relaxed to missing at random (MAR)

Outline

- 1 The problem
- 2 Conditional independent clustering**
- 3 Estimation
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp
- 6 Conclusion

High-dimensional today's data²



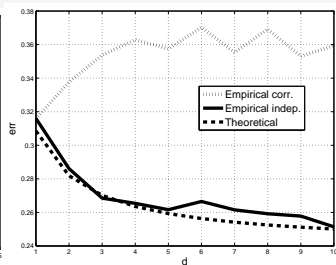
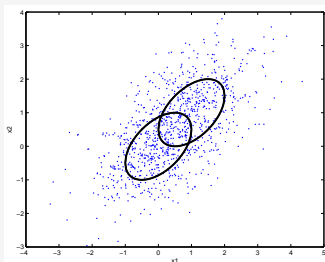
²S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, 29

HD clustering: blessing (1/2)

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

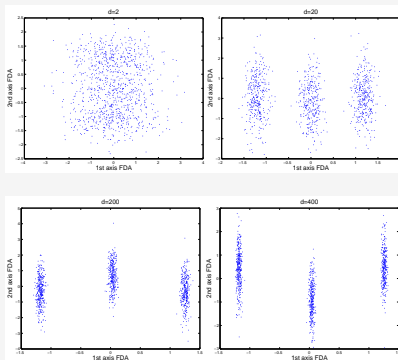
$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1|z_{11} = 1 \sim N_d(\mathbf{0}, \Sigma), \quad \mathbf{X}_1|z_{12} = 1 \sim N_d(\mathbf{1}, \Sigma)$$

- Each variable provides **equal** and **own** separation information
- Theoretical error decreases** when d grows: $\text{err}_{\text{theo}} = \Phi(-\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\Sigma^{-1}}/2)$
- Empirical error rate with the (true) **intra-correlated model worse** with d
- Empirical error rate with the (false) **intra-independent model better** with d !



HD clustering: blessing (2/2)

FDA



Neglect intra-dependency in HD clustering for better bias/variance trade-off^a

^aWhen variables convey no redundant cluster information; see conclusion

Mixed data: conditional independence everywhere

The aim is to combine continuous, categorical and integer data

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

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

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

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

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

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

Outline

- 1 The problem
- 2 Conditional independent clustering
- 3 Estimation**
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp
- 6 Conclusion

SEM algorithm

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

$$\ell(\theta; \mathbf{x}^O) = \ln p(\mathbf{x}^O; \theta)$$

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

Properties

- 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 $(\mathbf{x}^M, \mathbf{z})$

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

Properties

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

Outline

- 1 The problem
- 2 Conditional independent clustering
- 3 Estimation
- 4 Clustering with MixtComp**
- 5 Imputation with MixtComp
- 6 Conclusion

Prostate cancer data³

- **Individuals:** 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- **Variables:** $d = 12$ pre-trial variates were measured on each patient, composed by **eight continuous** variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histologic grade, serum prostatic acid phosphatase) and **four categorical** variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Some **missing data:** 62 missing values ($\approx 1\%$)

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

Questions

- How many clusters?
- Which partition?

³Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Create a free account in MixtComp⁴

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

It implements the mixed/missing data clustering in a software as a service (SaaS)

BigStat MixtComp HDPenReg MixAll BlockCluster Dev [Login](#) [Register](#)

Log in

Username
biernacki

Password

Log in

[Forgot Password?](#)

⁴See documentation at <https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp>

Two files to merge into a unique zip file

Variable descriptor file: descriptor.csv

probate_descriptor.csv - OpenOffice Calc

A	B	C	D	E	F	G	H	I	J	K	L	M
z_class	Age	Wt	PF	Hx	SBP	DBP	EKG	HG	SZ	SG	AP	BM
z	LaterClass	Gaussian_sjk	Gaussian_sjk	Categorical_sjk	Categorical_sjk	Gaussian_sjk	Gaussian_sjk	Categorical_sjk	Gaussian_sjk	Gaussian_sjk	Gaussian_sjk	Categorical_sjk

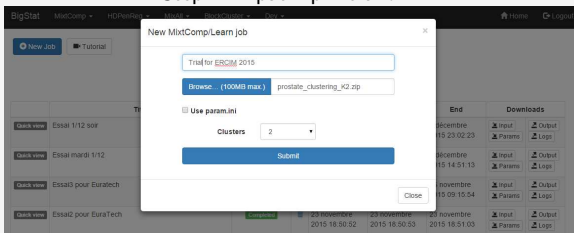
Data file: data.csv

data.csv - OpenOffice Calc

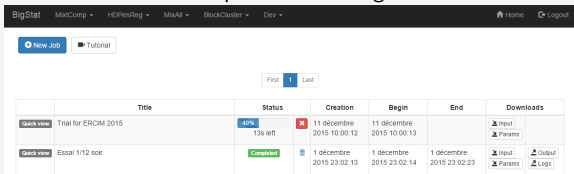
A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	z_class	Age	Wt	PF	Hx	SBP	DBP	EKG	HG	SZ	SG	AP	BM	
2	?	75	76	1	1	15	9	5	138	1.4142	8.1	0.986	1	
3	?	54	116	1	1	13	7	4	146	6.4807	?	1.9459	1	
4	?	69	102	1	2	14	8	5	134	1.7321	9.1	0.986	1	
5	?	75	94	2	2	14	7	2	176		2	8.2	1.972	1
6	?	67	99	1	1	17	10	1	134	5.831		8.1	6.094	1
7	?	71	98	1	1	19	10	1	151	3.1623	11.1	7.918	1	
8	?	75	100	1	1	14	10	2	130	3.6056		9.2	0.794	1
9	?	73	114	1	2	17	11	5	126	1.7321		9.1	7.918	1
10	?	60	110	1	1	12	8	1	146		2	10.1	9.459	1
11	?	78	107	1	2	13	8	6	130	4.5826		6.1	3.863	1
12	?	77	89	1	1	15	8	1	156	1.7321		8.1	7.918	1
13	?	74	105	1	2	18	14	1	136	2.4495		8.1	3.863	1
14	?	74	107	1	1	14	9	6	144	2.4495		9.1	0.986	1
15	?	55	112	1	2	16	9	5	139		2	9.2	3.026	1
16	?	73	88	1	1	19	10	5	120	3.873		10.1	7.918	1
17	?	87	81	2	2	17	12	3	134	1.7321		9.1	3.863	1
18	?	64	90	1	1	14	8	1	162	2.4495		9.1	9.459	1
19	?	79	104	1	1	13	8	2	150	2.2361		8.1	6.094	1
20	?	62	90	1	2	13	8	2	144	1.4142		9.1	9.459	1

Learn!

Step 1: input zip file and K






Step 2: it is running!

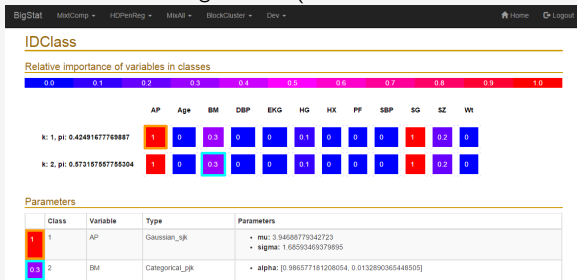


Output

Option 1: output zip file

<input type="checkbox"/> Nom	Modifié le	Type	Taille
 IDClass.html	23/11/2015 19:01	Fichier HTML	5 Ko
 log.txt	23/11/2015 19:01	Document texte	1 Ko
 output.RData	23/11/2015 19:01	R Workspace	54 Ko

Option 2: instant viewing clusters (variable-wise normalized entropy)



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

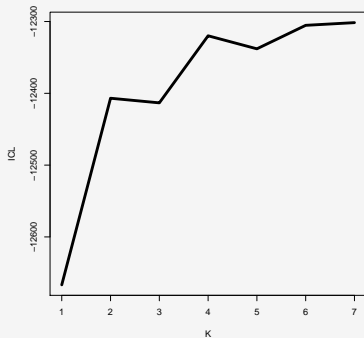
- **Strategy “mice⁵ + MixtComp”**: MixtComp on the dataset completed by mice

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

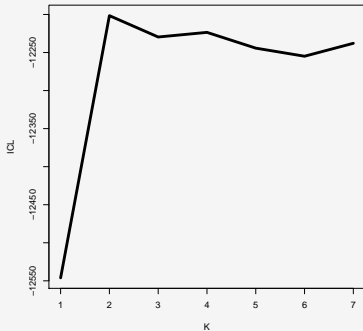
- **Strategy “full MixtComp”**: MixtComp on the observed (no completed) dataset

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

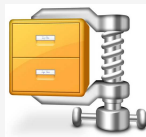
Strategy	continuous only	categorical only	mixed cont/cat
% misclassified	9.46	47.16	8.63

- 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



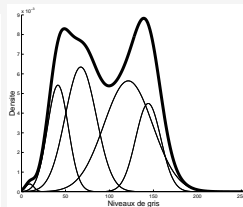
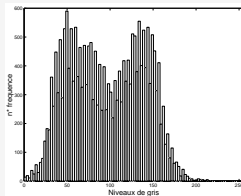
Then same output format as the **learn** functionality of MixtComp

Outline

- 1 The problem
- 2 Conditional independent clustering
- 3 Estimation
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp**
- 6 Conclusion

Mixture models as a extremely flexible family of distributions

Allow to estimate any distribution by increasing the number of components



Cancer dataset with more missing data

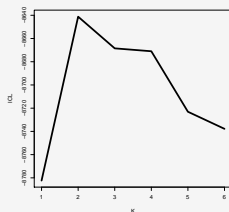
Add artificially $\approx 30\%$ missing data with a MCAR design

Then compare two strategies of imputation:

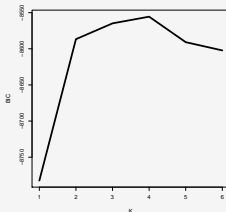
- Strategy “mice”: dataset completed by mice

```
> 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
 $\hat{K} = 4$

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 The problem
- 2 Conditional independent clustering
- 3 Estimation
- 4 Clustering with MixtComp
- 5 Imputation with MixtComp
- 6 Conclusion**

Present and future of MixtComp

■ Present

- Clustering and/or imputation for mixed/mixing/incertain data
- Current variables: continuous, categorical, integer
- Limit highly the preprocessing step: upload data as they are
- Software as a Service (SaaS) facility, nothing to install on the laptop
- Output: R objects and friendly/interactive graphical displays

■ Future

- Add other kinds of widespread variables: ordinal, ranks, functional, directional
- Add variable selection ability for tackle (very) high dimension: variable clustering?
- Improve gradually the server computing performance

