



Unifying Data Units and Models in (Co-)Clustering

C. Biernacki

Joint work with A. Lourme

CLADAG 2019 – 12th Scientific Meeting Classification and Data Analysis Group
Cassino (Italy), September 11 – 13, 2019



○○○○
○○○○○○○○
○○○
○○○○

○○○○
○○○○○
○○○○○

○○
○○○

Quizz!

$$y = \beta x^2 + e$$

- Is it a **linear** regression on co-variates (x^2)?
- Is it a **quadratic** regression on co-variates x ?

Both!

○○○○○
○○○○○○○○
○○○○
○○○○○

○○○○
○○○○○
○○○○○

○○
○○○

Take home message

Units are entirely interrelated with models

This part:

- Be aware that interpretation of (“classical”) models is **unit dependent**
- Models should even be revisited as a **couple units \times “classical” models**
- Opportunity for **cheap/wide/meaningful** enlarging of “classical” model families
- Focus on **model-based (co-)clustering** but larger potential impact

○○○○○
○○○○○○○○
○○○○
○○○○○

○○○○
○○○○○
○○○○○

○○
○○○

Outline

1 Introduction

2 Units in model-based clustering

- Scale units and parsimonious Gaussians
- Non scale units and Gaussians
- Class conditional units and Gaussians
- Units and Poissons

3 Units in model-based co-clustering

- Model for different kinds of data
- Units and Bernoulli
- Units and multinomial

4 Conclusion

- Summary
- Units and other distributions



General (model-based) statistical framework

■ Data:

- Whole data set composed by n **objects**, described by d **variables**

$$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \quad \text{with} \quad \mathbf{x}_i = (x_i^1, \dots, x_i^d) \in \mathbb{X}$$

- Each x_i value is provided with a **unit id**
- We note “**id**” since units are often user defined (a kind of canonical units)

■ Model:

- A pdf¹ family, indexed by $\mathbf{m} \in \mathbb{M}^2$

$$p_{\mathbf{m}} = \{ \cdot \in \mathbb{X} \mapsto p(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}} \}$$

- With $p(\cdot; \boldsymbol{\theta})$ a (parametric) pdf and $\Theta_{\mathbf{m}}$ a space where evolves this parameter

■ Target:

$$\widehat{\text{target}} = \mathbf{f}(\mathbf{x}, p_{\mathbf{m}})$$

Unit **id** is hidden everywhere and could have consequences on the target estimation!

¹probability density function

²Often, the index \mathbf{m} is confounded with the distribution family itself as a shortcut

○○○○○
 ○○○○○○○○
 ○○○○
 ○○○○

○○○○
 ○○○○○
 ○○○○○

○○
 ○○○
 ○○○

Changing the data units

- Principle of **data units transformation \mathbf{u}** :

$$\mathbf{u} : \mathbb{X} = \mathbb{X}^{\text{id}} \quad \longrightarrow \quad \mathbb{X}^{\mathbf{u}}$$

$$\mathbf{x} = \mathbf{x}^{\text{id}} = \text{id}(\mathbf{x}) \quad \longmapsto \quad \mathbf{x}^{\mathbf{u}} = \mathbf{u}(\mathbf{x})$$

- \mathbf{u} is a **bijective** mapping to preserve the whole data set information quantity
- We denote by \mathbf{u}^{-1} the reciprocal of \mathbf{u} , so $\mathbf{u}^{-1} \circ \mathbf{u} = \text{id}$
- Thus, id is only a particular unit \mathbf{u}
- Often a **meaningful** restriction³ on \mathbf{u} : it proceeds lines by lines and col. by col.

$$\mathbf{u}(\mathbf{x}) = (\mathbf{u}(\mathbf{x}_1), \dots, \mathbf{u}(\mathbf{x}_n)) \quad \text{with} \quad \mathbf{u}(\mathbf{x}_i) = (\mathbf{u}_1(\mathbf{x}_i^1), \dots, \mathbf{u}_d(\mathbf{x}_i^d))$$

- Advantage to respect the variable definition, transforming only its unit
- $\mathbf{u}(\mathbf{x}_i)$ means that \mathbf{u} applied to the data set \mathbf{x}_i , restricted to the single individual i
- \mathbf{u}_j corresponds to the specific (bijective) transformation unit associated to variable j

³Possibility to relax this restriction, including for instance linear transformations involved in PCA (principal component analysis). But the variable definition is no longer respected.



Revisiting units as a modelling component

- Explicitly exhibiting the “canonical” unit **id** in the model

$$P_{\mathbf{m}} = \{\cdot \in \mathbb{X} \mapsto p(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\} = \{\cdot \in \mathbb{X}^{\text{id}} \mapsto p(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\} = P_{\mathbf{m}}^{\text{id}}$$

- Thus the variable space and the probability measure are **embedded**
- As the **standard probability theory**: a couple (variable space, probability measure)!
- Changing **id** into **u**, while preserving **m**, is expected to produce a new modelling

$$P_{\mathbf{m}}^{\mathbf{u}} = \{\cdot \in \mathbb{X}^{\mathbf{u}} \mapsto p(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}.$$

A model should be systematically defined by a couple (\mathbf{u}, \mathbf{m}) , denoted by $P_{\mathbf{m}}^{\mathbf{u}}$



Interpretation and identifiability of p_m^u

- Standard probability theory (again): there exists a measure $\mathbf{u}^{-1}(\mathbf{m})$ s.t.⁴

$$\mathbf{u}^{-1}(\mathbf{m}) \in \{\mathbf{m}' \in \mathbb{M} : p_{\mathbf{m}'}^{\text{id}} = p_{\mathbf{m}}^u\}$$

- There exists **two alternative interpretations** of strictly the same model:
 - p_m^u : data measured with **unit u** arises from **measure m**;
 - $p_{\mathbf{u}^{-1}(\mathbf{m})}^{\text{id}}$: data measured with **unit id** arise from **measure $\mathbf{u}^{-1}(\mathbf{m})$**
- Two points of view:

Statistician

The model p_m^u is not identifiable over the couple (\mathbf{m}, \mathbf{u})

Practitioner

Freedom to choose the interpretation which is the most meaningful for her/him

⁴This set is usually restricted to a single element.

○○○○
○○○○○○○
○○○
○○○○

○○○
○○○○
○○○○
○○○○

○○
○○○

Opportunity for designing new models

Great opportunity to **build** easily numerous new **meaningful models** p_m^u !

- Just **combine** a standard model family $\{\mathbf{m}\}$ with a standard unit family $\{\mathbf{u}\}$
- New family can be huge! **Combinatorial problems** can occur. . .
- **Some model stability** can exist in some (specific) cases: $\mathbf{m} = \mathbf{u}^{-1}(\mathbf{m})$

○○○○○
 ○○○○○○○○
 ○○○○
 ○○○○

○○○○
 ○○○○○
 ○○○○

○○
 ○○○

Model selection

As any model, possible to choose between $p_{m_1}^{u_1}$ and $p_{m_2}^{u_2}$

However, caution when using likelihood-based model selection criteria (as BIC)

- **Prohibited** to compare m_1 in unit u_1 and m_2 in unit u_2
- But **allowed** after transforming in **identical unit id**
- Thus compare their equivalent expression: $p_{u_1^{-1}(m_1)}^{\text{id}}$ and $p_{u_2^{-1}(m_2)}^{\text{id}}$
- Example for abs. continuous x and differentiable u , the **density transform** in **id** is:

$$p_{u^{-1}(m)}^{\text{id}} = \{ \cdot \in \mathbb{X}^{\text{id}} \mapsto p(\mathbf{u}(\cdot); \theta) \times |\mathbf{J}^u(\cdot)| : \theta \in \Theta_m \}$$

with $\mathbf{J}^u(\cdot)$ the **Jacobian** associated to the transformation u



Focus on the clustering target

A current challenge is to enlarge model collection. . . and units could contribute to it!

- **Model:** mixture model \mathbf{m} of parameter $\theta = \{\pi_k, \alpha_k\}_{k=1}^g$

$$p_{\mathbf{m}}(\mathbf{x}; \theta) = \sum_{k=1}^g \pi_k p(\mathbf{x}; \alpha_k)$$

- g is the number of clusters
- Clusters correspond to a hidden partition $\mathbf{z} = (z_1, \dots, z_n)$, where $z_i \in \{1, \dots, g\}$
- $\pi_k = p(Z = k)$ and $p(\mathbf{x}; \alpha_k) = p(\mathbf{X} = \mathbf{x} | Z = k)$
- **Target:** estimate \mathbf{z} (and often g)
 - Estimate $\hat{\theta}_{\mathbf{m}}$ by maximum likelihood (typically)
 - Estimate \mathbf{z} by the MAP principle $\hat{z}_i = \arg \max_{k \in \{1, \dots, g\}} p(Z_i = k | \mathbf{X}_i = \mathbf{x}_i; \hat{\theta}_{\mathbf{m}})$
 - Estimate g by BIC or ICL criteria typically (maximum likelihood based criteria)



Outline

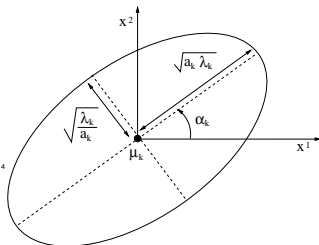
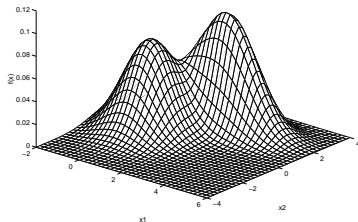
- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



14 spectral models on Σ_k

- $\mathbf{X} = \mathbb{R}^d$
- d -variate Gaussian model \mathbf{m} : $p_{\mathbf{m}}(\cdot; \alpha_k) = \mathcal{N}_d(\mu_k, \Sigma_k)$
- [Celeux & Govaert, 1995]⁵ propose the following **eigen decomposition**

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



⁵Celeux, G., and Govaert, G.. Gaussian parsimonious clustering models. Pattern Recognition, 28(5), 781–793 (1995).



Scale unit invariance

- Consider scale unit transformation $\mathbf{u}(\mathbf{x}) = \mathbf{D}\mathbf{x}$, with diagonal $\mathbf{D} \in \mathbb{R}^{d \times d}$
- Very **current transformation**: standard units (mm, cm), standardized units
- [Biernacki & Lourme, 2014] listed models where invariance holds (8 among 14)
 - The general model is invariant:

$$[\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k] = \mathbf{u}^{-1}([\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k])$$

- An example of not invariant model:

$$[\lambda_k \mathbf{S} \mathbf{\Lambda}_k \mathbf{S}'] \neq \mathbf{u}^{-1}([\lambda_k \mathbf{S} \mathbf{\Lambda}_k \mathbf{S}'])$$

- Do not forget to compare all models $\mathbf{m}' = \mathbf{u}^{-1}(\mathbf{m})$ in **unit id** for BIC / ICL validity
- Use the **Rmixmod** package (or any other **classical** package)



MASSICCC platform for the MIXMOD software

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

MASSICCC Help Login Sign up

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!



MixMod

Mixmod is a well-established software package for fitting a mixture model of multivariate Gaussian or multinomial probability distribution functions to a given data set. Cluster analysis will partition observations into groups ("clusters") while classification analysis will design a decision function from a learning data set to assign new data to groups a priori known.

[Read more about Mixmod](#)

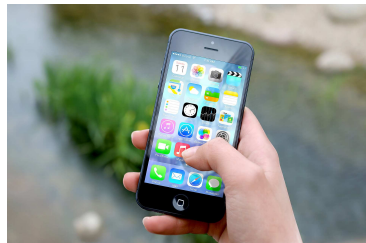




Illustration on the Old Faithful geyser data set

- All models are with free proportions (π_k)
- All ICL values are expressed with the initial unit $\text{id} = \text{min} \times \text{min}$
- We observe the **effect of unit on the ICL ranking** for some models
- **Cheap** opportunity to **enlarge** the model family!

family	$\text{id} = (\text{min}, \text{min})$		$\mathbf{u}^{\text{scale}_1} = (\text{sec}, \text{min})$		$\mathbf{u}^{\text{scale}_2} = (\text{stand}, \text{stand})$	
	\mathbf{m}	ICL^{id}	\mathbf{m}	ICL^{id}	\mathbf{m}	ICL^{id}
All mod.	$[\lambda_k \mathbf{S} \mathbf{\Lambda}_k \mathbf{S}']$	1 160.3	$[\lambda_k \mathbf{S} \mathbf{\Lambda}_k \mathbf{S}']$	1 158.7	$[\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k]$	1 160.3
General mod.	$[\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k]$	1 161.4	$[\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k]$	1 161.4	$[\lambda_k \mathbf{S}_k \mathbf{\Lambda}_k \mathbf{S}'_k]$	1 161.4



Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - **Non scale units and Gaussians**
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



Partitioning communes of Wallonia

- **Data:** $n = 262$ communes of Wallonia in terms of $d = 2$ fractals at a local level
 - 1st variable: fractal dimension of city **boundary** picture
 - 2nd variable: fractal dimension of city **surface** picture
- See more details in [Thomas *et al.*, 2008]⁶

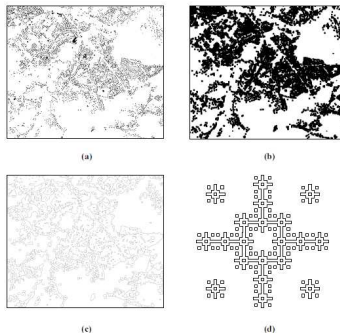


Figure 6: Extracting boundaries by dilation. Figure 6a shows the original urban pattern, Figure 6b the corresponding dilated structure (3 steps), Figure 6c the extracted boundary and Figure 6d a theoretical fractal with similar features to the observed fractal in (c)

⁶I. Thomas, P. Frankhauser and C. Biernacki (2008). The morphology of built-up landscapes in Wallonia (Belgium): a classification using fractal indices. *Landscape and Urban Planning*, 84, 99-115.



Results for Wallonia

- BIC retains $\mathbf{u} = (\text{exp}, \text{exp})$ and $\mathbf{m} = (\pi_k)[\lambda]$ (among **id**/log/exp and 14 spectral)
- Meaningful groups with $\mathbf{u} = (\text{exp}, \text{exp})$
- exp was a natural unit at the fractal level (“fractal dimension”)
- **exp also natural** since it corresponds to the “number of pixel pair comparisons”
- Somewhere, exp is quite related to the Manly transformation (see later)

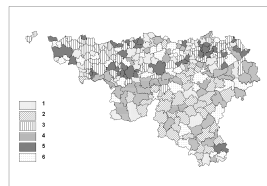


Partitioning overview for Wallonia

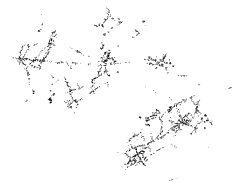
Table 3: Results of the fractal classification of communes

Cluster	$D_{\text{Surf-d0}}$	$D_{\text{bord-d5}}$	n	Three most representative communes	Typology
1	1.37	1.65	44	Brugelette, Heron, Nandrin	Peri-urban I and small cities
2	0.92	1.50	40	Lierneux, Havelange, Merbes-le-C.	Rural I: compact isolated hamlets
3	1.50	1.76	49	Pepinster, Saint-Georges, Blegny	Peri-urban II and eastern part (Hainaut)
4	1.11	1.59	47	Erquelines, Baelen, Rendux	Rural II: hamlets with a linear structure
5	1.68	1.70	40	Ottignies, Châtelet, Chaudfontaine	Urban (homogeneous, fully urbanised communes)
6	1.25	1.63	42	Gesves, Jalhay, Ciney	Rural III: rural communes with hamlets and one (small) city centre

n : number of communes in the class;



Wallonie communes clustering



Heron



Chaudfontaine



Prostate cancer data of [Biar & Green, 1980]⁷

- **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 “SZ”, index of tumour stage and histologic grade, serum prostatic acid phosphatase “AP”)
 - **Two ordinal** variables (performance rating, cardiovascular disease history)
 - **Two categorical** variables with various numbers of levels (electrocardiogram code, bone metastases)
- Some **missing data:** 62 missing values ($\approx 1\%$)

⁷Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488



Clustering with the MixtComp software [Biernacki et al., 2016]¹⁰

- Two historical units for performing the clustering task:
 - Raw units **id**: [McParland & Gormley, 2015]⁸
 - Transformed data **u**: since SZ and AP are skewed, [Jorgensen & Hunt, 1996]⁹ propose

$$\mathbf{u}_{SZ} = \sqrt{\cdot} \text{ and } \mathbf{u}_{AP} = \ln(\cdot)$$

- Model **m** in Mixtcomp: full mixed data $\mathbf{x} = (\mathbf{x}^{cont}, \mathbf{x}^{cat}, \mathbf{x}^{ordi}, \mathbf{x}^{int}, \mathbf{x}^{rank})$ (missing data are allowed also) are simply modeled by **inter conditional independence**

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

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

⁸McParland, D. and Gormley, I. C. (2015). Model based clustering for mixed data: clustmd. arXiv preprint arXiv:1511.01720.

⁹Jorgensen, M. and Hunt, L. (1996). Mixture model clustering of data sets with categorical and continuous variables. In Proceedings of the Conference ISIS, volume 96, pages 375–384.

¹⁰MixtComp is a clustering software developed by Biernacki C., Iovleff I. and Kubicki V. and freely available on the MASSICCC web platform <https://massiccc.lille.inria.fr/>



MASSICCC platform for the MIXTCOMP software

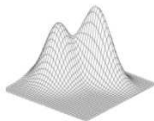
<https://massiccc.lille.inria.fr/>

MASSICCC Help Login Sign up

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.

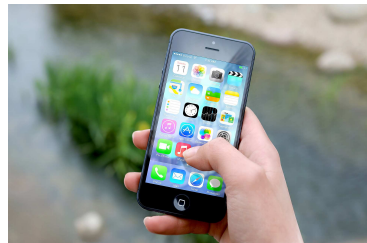
Developed by *Inria* TRY IT!



MixtComp

MixtComp takes mixture model analysis one step further and deals with mixed, missing or uncertain data which are common in today's data sets.

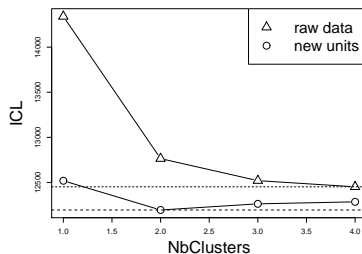
[Read more about MixtComp](#)





Results for the prostate cancer data set

- New units u_{SZ} and u_{AP} are selected by ICL
- New units allow to select **two groups** and provide a **lower error rate**



<i>clusters</i>	
1	2
287	5
52	162

Table : MixtComp model on raw units:
11% misclassified

<i>clusters</i>	
1	2
270	22
23	191

Table : MixtComp model on new units:
9% misclassified



Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - **Class conditional units and Gaussians**
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



Looking for conditional normality

- [Zhu & Melnykov, 2016]¹¹ transform units conditionally to classes for approaching class normality with the Manly transformation unit ($k = 1, \dots, g, j = 1, \dots, d$)

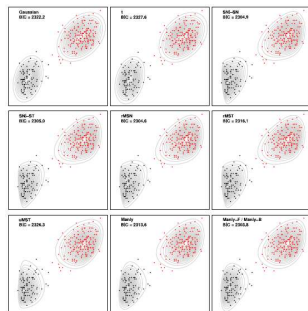
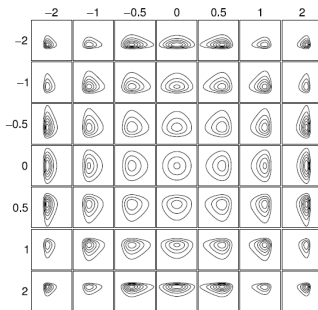
$$\mathbf{u}_\lambda = \{\mathbf{u}_{\lambda_{kj}}\} \quad \text{with} \quad \mathbf{u}_{\lambda_{kj}} = \begin{cases} \frac{\exp(\lambda_{kj}x^j) - 1}{\lambda_{kj}}, & \lambda_{kj} \neq 0 \\ x^j, & \lambda_{kj} = 0 \end{cases}$$

- Estimate parameters (θ, λ) by ml and the EM algorithm
- In fact choosing $\lambda_{kj} \in \{\mathbb{R}^+, \{0\}\}$ corresponds to a model and is performed by a forward and backward selection associated to a BIC criterion

¹¹Zhu, X. and Melnykov, V. (2016) Manly Transformation in Finite Mixture Modeling, accepted by Computational Statistics and Data Analysis.



Examples¹³



One bivariate component $\mathcal{N}_2(\mathbf{0}, \mathbf{I})$
Different $\lambda = (\lambda_1, \lambda_2)$ values

Old Faithful Geyser
[Azzalini & Bowman, 1990]¹²

¹²Azzalini, A., Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. J. Roy. Statist. Soc. Ser. C 39, 357–365.

¹³Figures from [Zhu & Melnykov, 2016]



Discussion on Manly units

- High flexibility for mixtures
- But low unit interpretation for two reasons
 - Manly transformation is a non-standard unit (?)
 - Unit transformation is class-dependent. . .
- Defend invariance of scale transformation of Manly as a desirable property. . .

2.4. Properties of Manly components

In this section, it is demonstrated that the proposed Manly components are invariant to the shifting and scaling of data points which is extremely desirable in the model-based clustering context as these operations should not lead to a different clustering result. For example, the relationship between temperature measured in degrees Celsius (C) and Fahrenheit (F) is given by $C = 5F/9 - 160/9$ and estimated partitions should be consistent under both scales.

. . . but it could be an opportunity to have no stability (provide new models!)



Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



Which units for count data?

- Count data: $x \in \mathbb{N}$
- Standard model \mathbf{m} is Poisson: $p(\cdot; \alpha_k) = \mathcal{P}(\lambda_k)$
- d -variate case $\mathbf{x} = (x^1, \dots, x^d) \in \mathbb{N}^d$ and conditional independence by variable
- Two standard unit transformations (by variable $j \in \{1, \dots, d\}$):
 - Shifted observations: $\mathbf{u}(x^j) = x^j - a_j$ with $a_j \in \mathbb{N}$
 - Scaled observations: $\mathbf{u}(x^j) = b_j x^j$ with $b_j \in \mathbb{N}^*$

Shifted example

- **id**: total number of educational years
- $\mathbf{u}_{\text{shift}}(\cdot) = (\cdot) - 8$: university number of educational years^a

^aEight is the number of years spent by English pupils in a secondary school.

Scaled example

- **id**: total number of educational years
- $\mathbf{u}_{\text{scaled}}(\cdot) = 2 \times (\cdot)$: total number of educational semesters



Medical data

- R dataset `rwm1984COUNT` of [Rao et al., 2007, p.221]¹⁴ and studied in [Hilbe, 2014]¹⁵
- $n = 3874$ patients that spent time into German hospitals during year 1984
- Patients are described through eleven mixed variables
- **m**: a MixtComp model combining Gaussian, Poisson and multinomial distributions

	<i>variables</i>	<i>type</i>	<i>model</i>
1	number of visits to doctor during year	count	Poisson
2	number of days in hospital	count	Poisson
3	educational level	categorical	multinomial
4	age	count	Poisson
5	outwork	binary	Bernoulli
6	gender	binary	Bernoulli
7	matrimonial status	binary	Bernoulli
8	kids	binary	Bernoulli
9	household yearly income	continous	Gaussian
10	years of education	count	Poisson
11	self employed	binary	Bernoulli

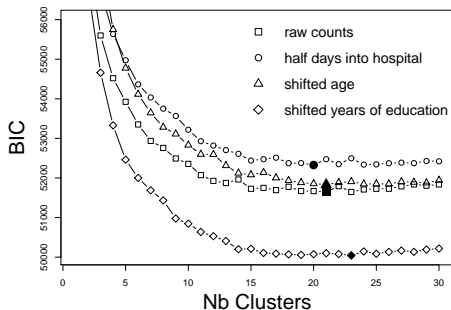
¹⁴Rao, C. R., Miller, J. P., and Rao, D. C. (2007). Handbook of statistics: epidemiology and medical statistics, volume 27. Elsevier.

¹⁵Hilbe, J. M. (2014). Modeling count data. Cambridge University Press.



Several units for count data

- **Four unit systems** are sequentially considered differing over the count data
 - $u_1 = \text{id}$: original unit
 - u_2 : the time spent into hospital is counted in half days instead of days
 - u_3 : the minimum of the age series is deduced from all ages leading to shifted ages
 - u_4 : the min. of years of edu. is deduced from the series leading to shifted years of edu.
- BIC selects 23 clusters obtained under **shifted years** of education





Specific transformation for RNA-seq data

- A sample of RNA-seq gene expressions arising from the rat count table of <http://bowtie-bio.sourceforge.net/recount/>
- 30000 genes described by 22 **counting** descriptors
- Remove genes with low expression (classical): 6173 genes finally
- Two different processes for dealing with data:
 - **Standard** [Rau *et al.*, 2015]¹⁶: $\mathbf{u} = \mathbf{id}$ and \mathbf{m} is Poisson mixture
 - **"RNA-seq unit"** [Gallopain *et al.*, 2015]¹⁷:

$$\mathbf{u}(\cdot) = \ln(\text{scaled normalization}(\cdot))$$

is a transformation being motivated by genetic considerations and \mathbf{m} is Gaussian mixture

- Experiment with 30 clusters (as in [Gallopain *et al.*, 2015])

<i>model</i>	<i>data</i>	<i>BIC</i>
Poisson	raw unit	2 615 654
Gaussian	transformed	909 190

¹⁶Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L. and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31 (9), 1420-1427.

¹⁷Gallopain, M., Rau, A., Celeux, G., and Jaffrézic, F. (2015). Transformation des données et comparaison de modèles pour la classification des données rna-seq. In 47èmes Journées de Statistique de la SFdS.

○○○○○
○○○○○○○○
○○○○
○○○○○

●○○○
○○○○○
○○○○○

○○
○○○

Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - **Model for different kinds of data**
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



Co-clustering framework

- It corresponds to the following **specific mixture model** \mathbf{m} [Govaert and Nadif, 2014]¹⁸:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z}, \mathbf{w})} \prod_{i,j} \pi_{z_i} \rho_{w_j} p(x_i^j; \boldsymbol{\alpha}_{z_i w_j})$$

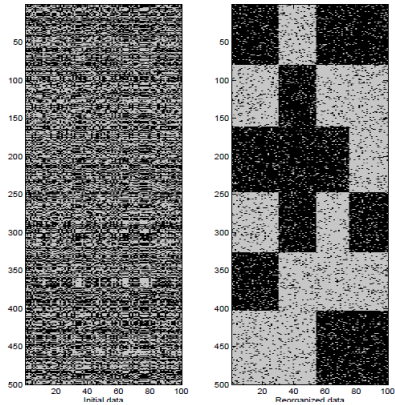
- \mathbf{z} : partition in g_r rows
- \mathbf{w} : partition in g_c columns
- $\mathbf{z} \perp \mathbf{w}$ and $x_i^j | (z_i, w_j) \perp x_{i'}^{j'} | (z_{i'}, w_{j'})$
- Distribution $p(\cdot; \boldsymbol{\alpha}_{z_i w_j})$ depends on the kind of data
 - Binary** data: $x_i^j \in \{0, 1\}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{B}(\boldsymbol{\alpha}_{kl})$
 - Categorical** data with m levels:
 - $\mathbf{x}_i^j = \{x_i^{jh}\} \in \{0, 1\}^m$ with $\sum_{h=1}^m x_i^{jh} = 1$ and $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{M}(\boldsymbol{\alpha}_{kl})$ with $\boldsymbol{\alpha}_{kl} = \{\alpha_k^{jh}\}$
 - Count** data: $x_i^j \in \mathbb{N}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{P}(\mu_k \nu_l \gamma_{kl})$
 - Continuous** data: $x_i^j \in \mathbb{R}$, $p(\cdot; \boldsymbol{\alpha}_{kl}) = \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$
- BlockCluster** [Bhatia et al., 2015]¹⁹ is an R package for co-clustering

¹⁸G. Govaert and M. Nadif (2014). Co-clustering: models, algorithms and applications. ISTE, Wiley. ISBN 978-1-84821-473-6.

¹⁹P. Bhatia, S. Iovleff, G. Govaert (2015). Blockcluster: An R Package for Model Based Co-Clustering. *Journal of Statistical Software*, in press.



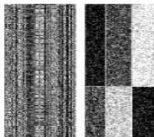
Binary illustration





MASSICCC platform for the BLOCKCLUSTER software

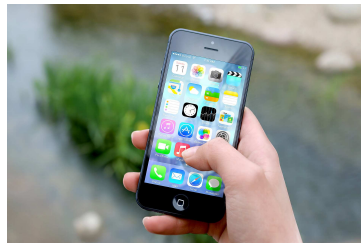
<https://massiccc.lille.inria.fr/>



BlockCluster

BlockCluster can estimate the parameters of co-clustering models for binary, contingency and continuous data. Simply put, when considering a set of data as rows and columns, BlockCluster will make simultaneous permutations of rows and columns in order to organise the data into homogenous blocks.

[Read more about BlockCluster](#)





Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - **Units and Bernoulli**
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions

```

○○○○○
○○○○○○○○
○○○○
○○○○○

```

```

○○○○
○●○○○
○○○○○

```

```

○○
○○○

```

SPAM E-mail Database²¹

- $n = 4601$ e-mails composed by 1813 “spams” and 2788 “good e-mails”
- $d = 48 + 6 = 54$ continuous descriptors²⁰
 - 48 percentages that a given **word** appears in an e-mail (“make”, “you’...)
 - 6 percentages that a given **char** appears in an e-mail (“;”, “\$”...)
- Transformation of continuous descriptors into **binary descriptors**

$$x_i^j = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

Two different units considered for variable $j \in \{1, \dots, 54\}$

- id_j : see the previous coding
- $\mathbf{u}_j(\cdot) = 1 - (\cdot)$: reverse the coding

$$\mathbf{u}_j(x_i^j) = \begin{cases} 0 & \text{if word/char } j \text{ appears in e-mail } i \\ 1 & \text{otherwise} \end{cases}$$

²⁰There are 3 other continuous descriptors we do not use

²¹<https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/>



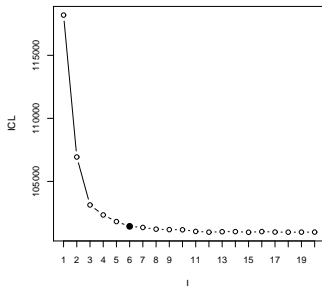
Select the whole coding $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$: pre-processing

- Impossible exhaustive algorithm to find the **best \mathbf{u} by ICL** (2^{54} possibilities)
- **Reduce the complexity** of the unit space, based on the following key idea:

Perform a clustering of the variables (columns only, no clusters in line)

“the more two variables have similar values (globally on lines)
the more a similar optimal unit transformation could be expected for both”

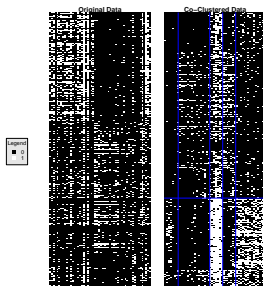
- A plateau of ICL at 6 clusters thus only $2^6 = 64$ possibilities (instead of $2^{54} \dots$)



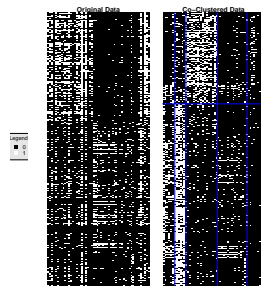


Select the whole coding $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$: co-clustering

- Fix $g_r = 2$ (two individual classes) and $g_c = 5$ (five variable classes)
- Use co-clustering in a **clustering aim**: just interested in indiv. classes (spams?)
- Exhaustive browse of 64 candidate units (those fixed by the pre-processing step)²²



initial unit id
ICL=92682.54
error rate=0.1984



best unit \mathbf{u}
ICL=92539.21
error rate=0.1999

²²Similar results if we select more clusters by pre-processing (e.g.: 14 col. clusters, thus $2^{14} = 16384$ candidates).



Result analysis of the e-mail database

- Just five variables (among 54 variables. . .) has a reversed coding in **u**
- Thus these five var. have **different coding than other ones** in their column class
- However, poor ICL increase with **u**

Conclusion for the e-mail database

- Here initial units **id** have a particular **meaning for the user**: do not change!
- In case of unit change, it becomes **essentially technic** (as Manly unit is)



Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions



Congressional Voting Records Data Set²⁴

- Votes for each of the $n = 435$ U.S. House of Representatives Congressmen
- Two classes: 267 democrats, 168 republicans
- $d = 16$ votes with $m = 3$ modalities [Schlimmer, 1987]²³:
 - “yea”: voted for, paired for, and announced for
 - “nay”: voted against, paired against, and announced against
 - “?”: voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known

- | | |
|--------------------------------------|--|
| 1. handicapped-infants | 9. mx-missile |
| 2. water-project-cost-sharing | 10. immigration |
| 3. adoption-of-the-budget-resolution | 11. synfuels-corporation-cutback |
| 4. physician-fee-freeze | 12. education-spending |
| 5. el-salvador-aid | 13. superfund-right-to-sue |
| 6. religious-groups-in-schools | 14. crime |
| 7. anti-satellite-test-ban | 15. duty-free-exports |
| 8. aid-to-nicaraguan-contras | 16. export-administration-act-south-africa |

²³Schlimmer, J. C. (1987). Concept acquisition through representational adjustment. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA.

²⁴<http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>

```

○○○○
○○○○○○○
○○○
○○○○

```

```

○○○○
○○○○○
○○●○○

```

```

○○
○○○

```

Allowed user meaningful recodings

- “yea” and “nea” are arbitrarily coded (**question dependent**), not “?”

- Example:

3. **adoption**-of-the-budget-resolution = “**yes**” \Leftrightarrow 3. **rejection**-of-the-budget-resolution = “**no**”

- However, “?” is **not question dependent**

Thus, two different units considered for variable $j \in \{1, \dots, 16\}$

- id_j :

$$x_i^j = \begin{cases} (1, 0, 0) & \text{if voted “yea” to vote } j \text{ by congressman } i \\ (0, 1, 0) & \text{if voted “nay” to vote } j \text{ by congressman } i \\ (0, 0, 1) & \text{if voted “?” to vote } j \text{ by congressman } i \end{cases}$$

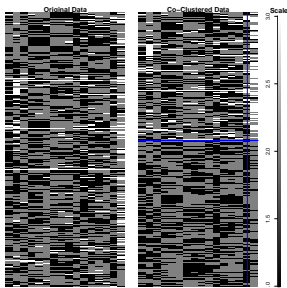
- $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$: reverse the coding **only for “yea” and “nea”**

$$\mathbf{u}_j(x_i^j) = \begin{cases} (0, 1, 0) & \text{if voted “yea” to vote } j \text{ by congressman } i \\ (1, 0, 0) & \text{if voted “nay” to vote } j \text{ by congressman } i \\ (0, 0, 1) & \text{if voted “?” to vote } j \text{ by congressman } i \end{cases}$$

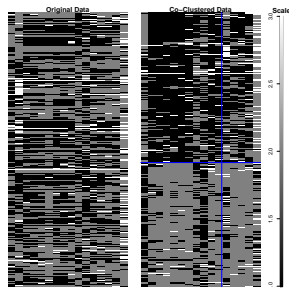


Select the whole coding $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$

- Fix $g_r = 2$ (two individual classes) and $g_c = 2$ (two variable classes)
- Use co-clustering in a **clustering aim**: just interested in political party
- Use a comprehensive algorithm to find the **best \mathbf{u}** by ICL ($2^{16} = 65536$ cases)



initial unit \mathbf{id}
 ICL=5916.13
 error rate=0.2850



best unit \mathbf{u}
 ICL=5458.156
 error rate=0.1034



Result analysis of the Congressional Voting Records Data Set

- Five variables have a reversed coding in \mathbf{u} :
 - 3. adoption-of-the-budget-resolution
 - 7. anti-satellite-test-ban
 - 9. aid-to-nicaraguan-contras
 - 10. mx-missile
 - 16. duty-free-exports
- Thus be aware to change the meaning of them when having a look at the figure!
- Significant **ICL and error rate improvements** with \mathbf{u}

Conclusion for the Congressional Voting Records

- Here initial units \mathbf{id} where arbitrary fixed: make sense to change!
- In addition, good improvement. . .

○○○○○
○○○○○○○○
○○○○
○○○○○

○○○○
○○○○○
○○○○○
○○○○○

●○
○○○

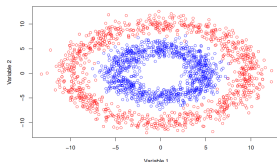
Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - **Summary**
 - Units and other distributions



Summary

- Be aware that interpretation of (“classical”) models is **unit dependent**
- Models should even be revisited as a **couple units** \times “classical” models
- Opportunity for **cheap/wide/meaningful** enlarging of “classical” model families



E.g.: easier to define a distribution with radial coordinates than with Cartesian ones

- But some units could be **user meaningful**, restricting this “technical enlarging”
- In counterpart, **combinatorial problems** may occur if the new family is huge

○○○○○
○○○○○○○○
○○○○
○○○○○

○○○○
○○○○○
○○○○○

○○
●○○

Outline

- 1 Introduction
- 2 Units in model-based clustering
 - Scale units and parsimonious Gaussians
 - Non scale units and Gaussians
 - Class conditional units and Gaussians
 - Units and Poissons
- 3 Units in model-based co-clustering
 - Model for different kinds of data
 - Units and Bernoulli
 - Units and multinomial
- 4 Conclusion
 - Summary
 - Units and other distributions

```

○○○○
○○○○○○○
○○○
○○○○

```

```

○○○○
○○○○○
○○○○○

```

```

○○
●●○

```

Units and other data types (and related distributions)

- **Ordinal** data $x \in \{\text{high grade, middle grade, low grade}\}$:
 - **id**: high grade > middle grade > low grade with “>” = greater in **strength** than
 - **u**: low grade > middle grade > high grade with “>” = greater in **weakness** than
 - Related distribution: see [Biernacki & Jacques, 2015]²⁵ and references therein
- **Ranking** data $x \in \{(\text{car,bike}), (\text{bike,car})\}$:
 - **id**: (car,bike) \Leftrightarrow car is preferred to bike, (bike,car) \Leftrightarrow bike is preferred to car
 - **u**: (car,bike) \Leftrightarrow bike is preferred to car, (bike,car) \Leftrightarrow car is preferred to bike
 - Related distribution: see [Jacques & Biernacki, 2014]²⁶ and references therein
- **Other**: directional data...

²⁵C. Biernacki and J. Jacques (2015). Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. Statistics and Computing, in press.

²⁶J. Jacques & C. Biernacki (2014). Model-based clustering for multivariate partial ranking data. Journal of Statistical and Planning Inference, 149, 201–217.

○○○○
○○○○○○○
○○○
○○○○

○○○
○○○○
○○○○

○○
○○●

Thank you for you attention