

# Contribution to model-based clustering of heterogeneous data


Vincent Vandewalle


Université de Lille, ULR 2694 & Inria Lille, Modal team  
École doctorale des sciences pour l'ingénieur


HDR defense  
Thursday, January 7, 2021



1. Introduction
2. Contribution to model-based clustering
  - General presentation
  - Focus on multiple partition clustering
3. Contribution to some applications
4. Research project

2003-2006:  Agronomy engineer diploma, AgroParistech, Paris


2005-2006:  M2 research in probability and statistics, University Paris XI, Orsay


2006-2009:  Ph.D. thesis in statistics, Lille University, Inria grant  
Title: Estimation and selection in semi-supervised classification

Supervisors: Christophe Biernacki and Gilles Celeux

Reviewers: Didier Chauveau and Jean-Jacques Daudin

President: Gérard Govaert

2009-2010:  Temporary assistant professor, Lille University, IUT C, STID department

2010-today:  Assistant professor, Lille University, IUT C, STID department

Affiliations:  ULR 2694: METRICS,

 Inria Lille: Project-Team MODAL

2012-2015: Chief of STID departement

2016-2017: CRCT (leave for research)

2020-2021: Delegation Inria

## Ph.D. theses supervision

- 2011-2014 Matthieu Marbac-Lourdelle, *Mixture models for clustering of categorical and mixed data* (with CB)
- 2016-2019 Adrien Ehrhardt, *Predictive models for big and biased data: application to improving credit scoring* (with CB and PH), CIFRE grant with CACF company
- 2019-2022 Wilfried Heyse, *Taking into account temporal structure in high throughput proteomic data* (with CBa and GM)
- 2020-2023 Axel Potier, *Sales predictions by grouping low turnover products* (with CB and MM), CIFRE grant with ADEO company

## Other supervisions

- six M2 internships: MM, KN, HC, TM, SI, WH
- two M1 internships: SZ, CF
- one post-doctoral researcher: Florent Dewez
- several engineers

## Participation to research projects

- 2013-2017 ANR project ClinMine: Optimization of patient care at hospital
- 2016-2018 CloHe project of Mastodons challenge of CNRS “Big data and data quality”
- 2018-2020 European project PERF-AI: Enhance Aircraft Performance and Optimisation through utilisation of Artificial Intelligence
- 2018-2022 ANR project Smiles: Statistical Modeling and Inference for unsupervised Learning at Large-Scale

## Participation to research contracts with companies

- 2011 Natural security: Hazard computation for a new payment technology
- 2014 Rouge gorge: Clustering clients
- 2014-2015 Auchan: Variable selection in regression
- 2015-2016 Cyland: Sale prediction







## Models are useful

- 💡 for clustering of heterogeneous data
- 💡 for allowing for several clustering viewpoints
- 💡 formalize the problem for applied issues







## Statistical toolbox

- 🔧 Statistical modeling
- 🔧 Model-based clustering tools: latent variable models, maximum likelihood estimation, EM algorithm, ...
- 🔧 Model choice: BIC, integrated likelihood, ICL, ...
- 🔧 Bayesian statistics: modeling, MCMC algorithms (Gibbs, MH), ...






## Contribution to model-based clustering (Part I)

- Chap. 3 Clustering of categorical and mixed data (3 : *JoC*, *ADAC*, *Comm. Statist. Theory Methods*, MM, CB)
- Chap. 4 Clustering of functional data (2 : *IRBM*, *Book Chapter*, 1 , CP, SD, QG)
- Chap. 5 Multiple partition clustering (2 : *CSDA*, *Mathematics*, MM)
- Chap. 6 General issues: label switching, missing data, visualization (3 : *AIP Conference Proceedings*, *Neurocomputing*, *JoC*, 1 , CB, MM, EE, AL)

## Contribution to model-based clustering (Part I)

- Chap. 3 Clustering of categorical and mixed data (3 : *JoC*, *ADAC*, *Comm. Statist. Theory Methods*, MM, CB)
- Chap. 4 Clustering of functional data (2 : *IRBM*, *Book Chapter*, 1 , CP, SD, QG)
- Chap. 5 Multiple partition clustering (2 : *CSDA*, *Mathematics*, MM)
- Chap. 6 General issues: label switching, missing data, visualization (3 : *AIP Conference Proceedings*, *Neurocomputing*, *JoC*, 1 , CB, MM, EE, AL)

## Contribution to some applications (Part II)

- Chap. 7 Usability study (1 : *BMC Med. Res. Methodol.*, 1 , AC, BD)
- Chap. 8 Artificial intelligence for aviation (1 : *DCE*, 1 , FD, BG, AT)
- Chap. 9 Credit scoring (2 , AE, CB, PH)



1. Introduction
2. Contribution to model-based clustering
  - General presentation
  - Focus on multiple partition clustering
3. Contribution to some applications
4. Research project

# Clustering setting

## Data

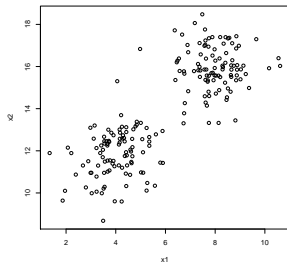
$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  an  $n$ -i.i.d sample, with  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ .

individuals \ variables	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_d$
$\mathbf{x}_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1d}$
$\mathbf{x}_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2d}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\mathbf{x}_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nd}$

## Goal

Clustering data in  $g$  clusters:

- Partition data in  $g$  homogeneous sub-populations
- Summarize the data
- Interpretation of each cluster
- Further use: predictive modeling, decision making, ...



# Clustering setting

## Data

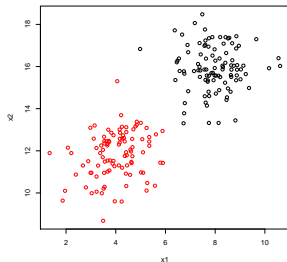
$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  an  $n$ -i.i.d sample, with  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ .

individuals \ variables	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_d$
$\mathbf{x}_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1d}$
$\mathbf{x}_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2d}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\mathbf{x}_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nd}$

## Goal

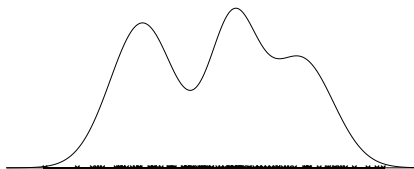
Clustering data in  $g$  clusters:

- Partition data in  $g$  homogeneous sub-populations
- Summarize the data
- Interpretation of each cluster
- Further use: predictive modeling, decision making, ...



## Motivation for model-based clustering

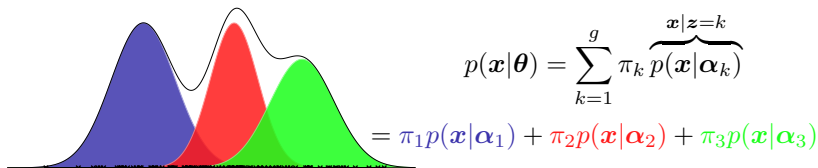
- A unique criterion to optimize: the likelihood
- Model selection: AIC, BIC, ICL, ...
- Handle different kinds of variables through the model
- Missing data naturally taken into account: EM algorithm



$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^g \pi_k \overbrace{p(\mathbf{x}|\boldsymbol{\alpha}_k)}^{\mathbf{x}|z=k}$$

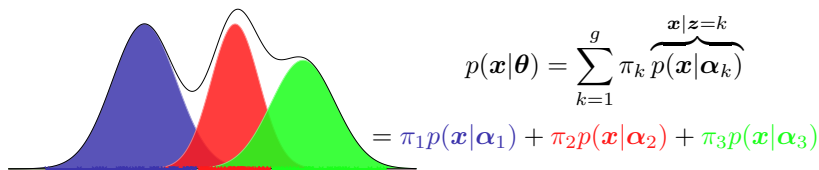
## Motivation for model-based clustering

- A unique criterion to optimize: the likelihood
- Model selection: AIC, BIC, ICL, ...
- Handle different kinds of variables through the model
- Missing data naturally taken into account: EM algorithm



## Motivation for model-based clustering

- A unique criterion to optimize: the likelihood
- Model selection: AIC, BIC, ICL, ...
- Handle different kinds of variables through the model
- Missing data naturally taken into account: EM algorithm



Estimated parameter  $\hat{\boldsymbol{\theta}} \Rightarrow$  Posterior class membership  $z|\mathbf{x}, \hat{\boldsymbol{\theta}}$   
 $\Rightarrow$  Estimated cluster  $\hat{z}$

- 🔒 Need for models on  $p(\mathbf{x}|\boldsymbol{\alpha}_k)$  for various kinds of data
- 🔒 Several interesting clustering
- 🔒 Missing data
- 🔒 Bayesian parameters estimation
- 🔒 Generic visualization

# Clustering of categorical and mixed data (1/2)

$$\text{Dependency per block: } p(\mathbf{x}|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) = \prod_{b=1}^{B_k} p(\mathbf{x}^{\{kb\}}|\boldsymbol{\theta}_{kb})$$

- 📍 Interpret intra-class dependency
- 📍 Relevant number of clusters



# Clustering of categorical and mixed data (1/2)

$$\text{Dependency per block: } p(\mathbf{x}|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) = \prod_{b=1}^{B_k} p(\mathbf{x}^{\{kb\}}|\boldsymbol{\theta}_{kb})$$

- 📍 Interpret intra-class dependency
- 📍 Relevant number of clusters

Mixture of intermediate dependency (Marbac, Biernacki, and Vandewalle, 2015)

$$p(\mathbf{x}^{\{kb\}}|\boldsymbol{\theta}_{kb}) = (1 - \rho_{kb}) \overbrace{p(\mathbf{x}^{\{kb\}}|\boldsymbol{\alpha}_{kb})}^{\text{independence}} + \rho_{kb} \overbrace{p(\mathbf{x}^{\{kb\}}|\boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})}^{\text{maximum dependency}}$$

- 🔒 Maximum dependency location
- 🔒 MH, GEM algorithm
- 🔒 Block structure
- 🔒 Gibbs sampler, BIC

# Clustering of categorical and mixed data (1/2)

$$\text{Dependency per block: } p(\mathbf{x}|\boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) = \prod_{b=1}^{B_k} p(\mathbf{x}^{\{kb\}}|\boldsymbol{\theta}_{kb})$$

- 📍 Interpret intra-class dependency
- 📍 Relevant number of clusters

Mixture of intermediate dependency (Marbac, Biernacki, and Vandewalle, 2015)

$$p(\mathbf{x}^{\{kb\}}|\boldsymbol{\theta}_{kb}) = (1 - \rho_{kb}) \overbrace{p(\mathbf{x}^{\{kb\}}|\boldsymbol{\alpha}_{kb})}^{\text{independence}} + \rho_{kb} \overbrace{p(\mathbf{x}^{\{kb\}}|\boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})}^{\text{maximum dependency}}$$

- 🔒 Maximum dependency location
- 🔒 MH, GEM algorithm
- 🔒 Block structure
- 🔒 Gibbs sampler, BIC

Dependency per modes (Marbac, Biernacki, and Vandewalle, 2016)

$$\mathbf{x}^{\{kb\}} \Leftrightarrow \tilde{x}^b \text{ (univariate)} + \text{distribution per modes on } \tilde{x}^b|\boldsymbol{\theta}_{kb} \quad \text{▮...}$$

- 🔒 Modes number
- 🔒 Integrated complete likelihood
- 🔒 Block structure
- 🔒 Gibbs sampler

# Clustering of categorical and mixed data (2/2)

Dependency per Gaussian copula  
(Marbac, Biernacki, and Vandewalle, 2017)

$$P(\mathbf{x}|\alpha_k) = \Phi_\epsilon(\Phi_1^{-1}(u_k^1), \dots, \Phi_1^{-1}(u_k^e) | \mathbf{0}, \Gamma_k) \text{ with } u_k^j = P(x^j | \beta_{k,j})$$

📍 Cluster continuous, discrete, ordinal data

📍 PCA-type component-based visualization

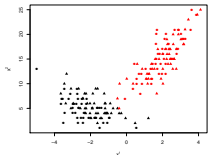
Computational challenge

🔒 Estimation of  $\Gamma_k$  and  $\beta_{k,j}$

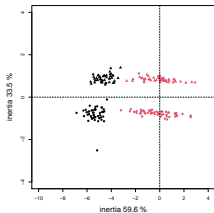
🔒 Full-Bayesian setting

🔒 Multivariate integrals

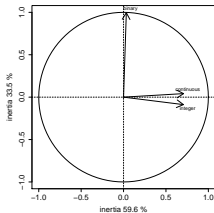
🔒 Metropolis-within-Gibbs



(a) Initial data



(b) Visualization component 2

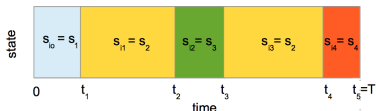


(c) Correlation component 2

# Clustering of functional data

Categorical functional data  
(Dhaenens et al., 2018)

- 🔒 Visualization
- 🔒 Different paths lengths

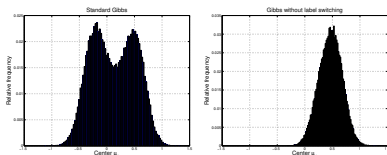


- 🔒 Functional MCA
- 🔒 Markov model on  $p(\mathbf{x}|\boldsymbol{\alpha}_k)$



## Label switching (Biernacki and Vandewalle, 2011)

🔒  $p(\boldsymbol{\theta}|\mathbf{x}) = p(\sigma(\boldsymbol{\theta})|\mathbf{x}), \forall \sigma \in \mathcal{P}_g$



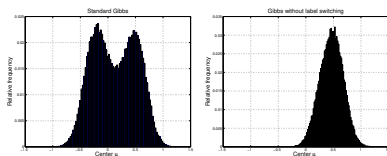
🔒  $p(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{g!} \sum_{h=1}^{g!} p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$   
 $\Rightarrow p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_1)$  unswitched posterior distribution

🔒 Choice of  $\mathcal{Z}_1^{MAP}$

🔒 Adapted Gibbs sampler

## Label switching (Biernacki and Vandewalle, 2011)

🔒  $p(\boldsymbol{\theta}|\mathbf{x}) = p(\sigma(\boldsymbol{\theta})|\mathbf{x}), \forall \sigma \in \mathcal{P}_g$



🔒  $p(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{g!} \sum_{h=1}^{g!} p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_h)$   
 $\Rightarrow p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{Z}_1)$  unswitched posterior distribution

🔒 Choice of  $\mathcal{Z}_1^{MAP}$

🔒 Adapted Gibbs sampler

## Missing data (Eirola et al., 2014; Vandewalle and Biernacki, 2015)

🔒 Compute distances with missing data:  $\mathbb{E}[\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 | \mathbf{x}_i^o, \mathbf{x}_{i'}^o]$

🔒 Slow degeneracy of the EM algorithm

🔒 Gaussian mixture on  $\mathbf{x}$  + adapted EM algorithm

🔒 Add constraints on the partition space  $\mathcal{Z}$

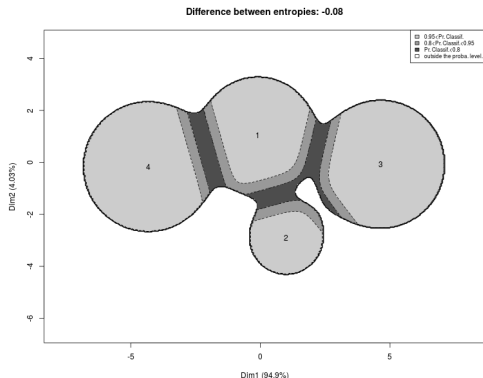
# Generic contributions to model-based clustering (2/2)

## Visualization in mixture (Biernacki, Marbac, and Vandewalle, 2020)

- 🔒 Depends on the nature of  $\mathbf{x}$
- 🔒 Control the arrival distribution family: constrained spherical Gaussian
- 🔒 Does not focus on clustering
- 🔒 Preserve posterior class membership distribution

## Congressmen data

- $n = 435$  congressmen
- $d = 16$  categorical variables with 3 levels
- Mixture of product of multinomial with  $g = 4$





# Multiple partition clustering overview (1/2)

## Limitation of standard clustering

- 🔒 Results highly depend of the chosen variables
- 🔒 Only dominant clustering viewpoint
- 🔒 Continuous variables dominate categorical variables
- 🔒 Class variable over-simplistic for predictive models
- 🔒 Large number of clusters needed

## Question

- ❓ Each variable comes with its own partition: how to find a trade-off between all these partitions?

## Multiple partition clustering: several cluster variables

- 💡 Several clustering viewpoints with respect to different groups/linear combinations of variables
- 💡 Find hidden clustering
- 💡 Summary of the data by several categorical variables

## State of the art in multiple partition clustering

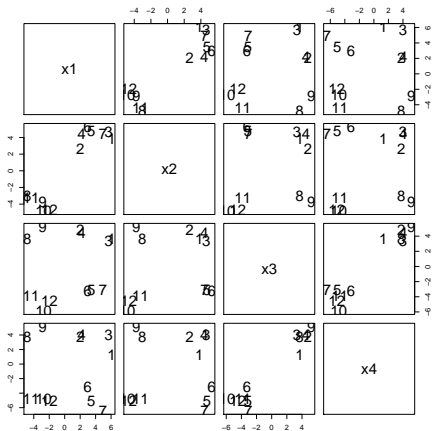
- 📖 Galimberti and Soffritti (2007): First multiple Gaussian mixture model
- 📖 Galimberti, Manisi, and Soffritti (2018): Refinement considering different roles for variables
- 📖 Poon et al. (2013): tree dependence structure between class variables
- 📖 Attias (1999): independent factor analysis

## Proposal

- 💡 Marbac and Vandewalle (2019): independent blocks of variables with conditional independence assumption in each block (mixed-data framework)
- 💡 Vandewalle (2020): several classifying linear combinations of variables (continuous framework)

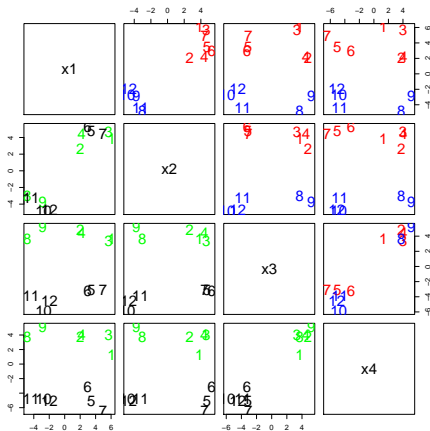
# Multiple partition by blocks of variables (1/5)

	$x_1$	$x_2$	$x_3$	$x_4$	$z_1$	$z_2$
$x_1$	6.1	3.8	3.7	1.3	?	?
$x_2$	2.0	2.7	4.8	3.8	?	?
$x_3$	5.7	4.7	3.3	4.0	?	?
$x_4$	2.2	4.4	4.5	4.1	?	?
$x_5$	3.4	4.7	-3.1	-5.0	?	?
$x_6$	3.0	5.3	-3.3	-3.1	?	?
$x_7$	4.9	4.5	-3.1	-6.5	?	?
$x_8$	-4.8	-3.0	3.7	3.8	?	?
$x_9$	-2.9	-3.7	5.2	5.1	?	?
$x_{10}$	-2.7	-4.8	-5.9	-4.8	?	?
$x_{11}$	-4.3	-3.2	-3.8	-4.7	?	?
$x_{12}$	-2.0	-4.6	-4.5	-5.1	?	?



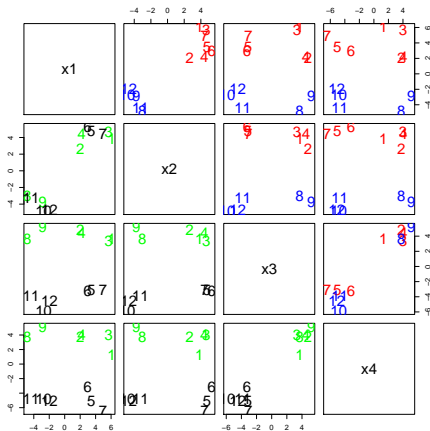
# Multiple partition by blocks of variables (1/5)

	$x_1$	$x_2$	$x_3$	$x_4$	$z_1$	$z_2$
$x_1$	6.1	3.8	3.7	1.3	1	1
$x_2$	2.0	2.7	4.8	3.8	1	1
$x_3$	5.7	4.7	3.3	4.0	1	1
$x_4$	2.2	4.4	4.5	4.1	1	1
$x_5$	3.4	4.7	-3.1	-5.0	1	2
$x_6$	3.0	5.3	-3.3	-3.1	1	2
$x_7$	4.9	4.5	-3.1	-6.5	1	2
$x_8$	-4.8	-3.0	3.7	3.8	2	1
$x_9$	-2.9	-3.7	5.2	5.1	2	1
$x_{10}$	-2.7	-4.8	-5.9	-4.8	2	2
$x_{11}$	-4.3	-3.2	-3.8	-4.7	2	2
$x_{12}$	-2.0	-4.6	-4.5	-5.1	2	2



# Multiple partition by blocks of variables (1/5)

	$x_1$	$x_2$	$x_3$	$x_4$	$z_1$	$z_2$
$x_1$	6.1	3.8	3.7	1.3	1	1
$x_2$	2.0	2.7	4.8	3.8	1	1
$x_3$	5.7	4.7	3.3	4.0	1	1
$x_4$	2.2	4.4	4.5	4.1	1	1
$x_5$	3.4	4.7	-3.1	-5.0	1	2
$x_6$	3.0	5.3	-3.3	-3.1	1	2
$x_7$	4.9	4.5	-3.1	-6.5	1	2
$x_8$	-4.8	-3.0	3.7	3.8	2	1
$x_9$	-2.9	-3.7	5.2	5.1	2	1
$x_{10}$	-2.7	-4.8	-5.9	-4.8	2	2
$x_{11}$	-4.3	-3.2	-3.8	-4.7	2	2
$x_{12}$	-2.0	-4.6	-4.5	-5.1	2	2



## Model assumptions

- 1  $B$  independent blocks of variables
- 2 Each block  $b$  follows a  $g_b$  mixture with class conditional independence

## Probability distribution function of $x_i$

$$p(\mathbf{x}_i | \mathbf{m}, \theta) = \prod_{b=1}^B p(\mathbf{x}_{i\{b\}} | \mathbf{m}, \theta) \text{ with } p(\mathbf{x}_{i\{b\}} | \mathbf{m}, \theta) = \sum_{k=1}^{g_b} \pi_{bk} \prod_{j \in \Omega_b} p(x_{ij} | \alpha_{jk}),$$

## Properties

- Permits **variable selection** (Marbac and Sedki, 2017) and **multiple partitions** explained by **subsets of variables** (variables classification) in an **heterogeneous** data setting
- Sparse model: low number of parameters
- Better model search expected than in Galimberti, Manisi, and Soffritti (2018)

## Model collection $\mathcal{M}$

- $\omega = (\omega_j; j = 1, \dots, d)$  the repartition of the variables in blocks;  $\omega_j = b$  if variable  $j$  belongs to block  $b$ .
- $\mathbf{m} = (g_1, \dots, g_B, \omega)$  defines the model

$$\mathcal{M} = \{\mathbf{m} : \omega_j \leq B_{\max} \text{ and } g_b \leq g_{\max}; j = 1, \dots, d; b = 1, \dots, B_{\max}\}$$

🔒  $|\mathcal{M}|$  large!

# Multiple partition by blocks of variables (3/5)

Tractable choice of  $\omega$  for  $B$  and  $(g_1, \dots, g_B)$  fixed  $\Leftrightarrow$  choice of  $\mathbf{m}$

Additive effect of the variables on the completed log-likelihood

$$\ell(\boldsymbol{\theta}_{\mathbf{m}} | \mathbf{m}, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^B \ln p(\mathbf{z}_b | \boldsymbol{\pi}_b) + \sum_{j=1}^d \ln p(\mathbf{x}_j | \mathbf{z}_{\omega_j}, \boldsymbol{\alpha}_j)$$

# Multiple partition by blocks of variables (3/5)

Tractable choice of  $\omega$  for  $B$  and  $(g_1, \dots, g_B)$  fixed  $\Leftrightarrow$  choice of  $m$

Additive effect of the variables on the completed log-likelihood

$$\ell(\theta_m | m, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^B \ln p(\mathbf{z}_b | \pi_b) + \sum_{j=1}^d \ln p(\mathbf{x}_j | \mathbf{z}_{\omega_j}, \alpha_j)$$

$\Rightarrow$  Modified EM algorithm increasing the BIC: add **affectation step** of each variable **individually** to the most accurate block

$$\text{BIC}(m) = \max_{\theta_m} \ell_{pen}(\theta_m | m, \mathbf{x}) \text{ with } \ell_{pen}(\theta_m | m, \mathbf{x}) = \ell(\theta_m | m, \mathbf{x}) - \frac{\nu m}{2} \ln n.$$

$$m^* = \operatorname{argmax}_m \text{BIC}(m) \Leftrightarrow (m^*, \hat{\theta}_{m^*}) = \operatorname{argmax}_{(m, \theta_m)} \ell_{pen}(\theta_m | m, \mathbf{x})$$

$\Rightarrow$  Alternate optimization between  $m$  and  $\theta_m$



# Multiple partition by blocks of variables (3/5)

Tractable choice of  $\omega$  for  $B$  and  $(g_1, \dots, g_B)$  fixed  $\Leftrightarrow$  choice of  $m$

Additive effect of the variables on the completed log-likelihood

$$\ell(\theta_m | m, \mathbf{x}, \mathbf{z}) = \sum_{b=1}^B \ln p(\mathbf{z}_b | \pi_b) + \sum_{j=1}^d \ln p(\mathbf{x}_j | \mathbf{z}_{\omega_j}, \alpha_j)$$

$\Rightarrow$  Modified EM algorithm increasing the BIC: add **affectation step** of each variable **individually** to the most accurate block

$$\text{BIC}(m) = \max_{\theta_m} \ell_{pen}(\theta_m | m, \mathbf{x}) \text{ with } \ell_{pen}(\theta_m | m, \mathbf{x}) = \ell(\theta_m | m, \mathbf{x}) - \frac{\nu m}{2} \ln n.$$

$$m^* = \operatorname{argmax}_m \text{BIC}(m) \Leftrightarrow (m^*, \hat{\theta}_{m^*}) = \operatorname{argmax}_{(m, \theta_m)} \ell_{pen}(\theta_m | m, \mathbf{x})$$

$\Rightarrow$  Alternate optimization between  $m$  and  $\theta_m$

$\Rightarrow$  Also possible in a fully Bayesian framework for the MICL

# Multiple partition by blocks of variables (4/5)

Application on wine data (1/2)

Data (available in the package pgmm)

- 27 chemical and physical properties of three types of Italian wines: Barolo, Grignolino, Barbera
- Data collected during the time period of 1970–1979

Models selected by BIC

$B$	BIC	Time	Block	G	ARI <sup>1</sup>
1	-6025.00	30	1	4	0.78
2	-5947.88	280	1	3	0.87
			2	4	0.16
3	-5921.42	1590	1	4	0.74
			2	4	0.20
			3	2	0.02
4	-5918.06	6065	1	4	0.75
			2	2	0.21
			3	3	0.02
			4	2	0.00

<sup>1</sup>ARI between the estimated partition and the type of wine

# Multiple partition by blocks of variables (5/5)

Application on wine data (2/2)

Block 1: the type of wines (ARI=0.75)

19 variables: Alcohol, Sugar-free Extract, Tartaric Acid, Uronic Acids, Alcalinity of Ash, Calcium, Magnesium, Phosphate, Total Phenols, Flavanoids, Non-flavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of Diluted Wines, OD280/OD315 of Flavanoids, Glycerol, 2-3-Butanediol, Proline

	Barolo	Grignolino	Barbera
Class 1	0	45	0
Class 2	0	5	48
Class 3	58	1	0
Class 4	1	20	0

Block 2: year of production

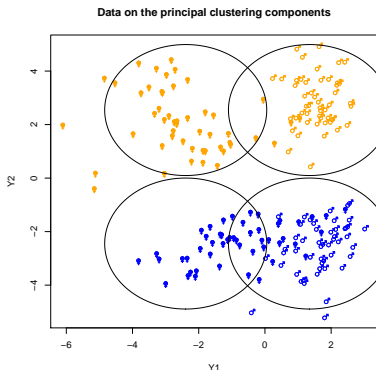
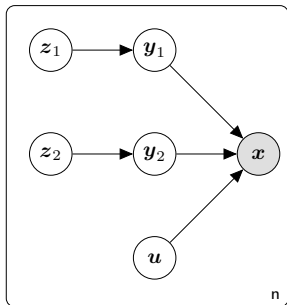
4 variables: Fixed Acidity, Malic Acid, pH, Total Nitrogen

	Year									
	1970	1971	1972	1973	1974	1975	1976	1978	1979	
Class 1	8	25	4	27	31	3	1	0	0	
Class 2	1	3	3	2	14	6	16	29	5	

# Multiple partition clustering subspaces (Vandewalle, 2020)

- Combine visualization and clustering
- Consider several partitions

- Model proposal
- GEM estimation algorithm



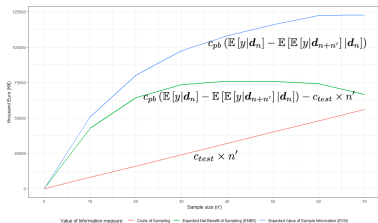
$$\mathbf{x} = A \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{u} \end{pmatrix} \stackrel{A \text{ nonsingular}}{\iff} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{u} \end{pmatrix} = A^{-1} \mathbf{x}$$

1. Introduction
2. Contribution to model-based clustering
  - General presentation
  - Focus on multiple partition clustering
3. Contribution to some applications
4. Research project

# Usability study (Vandewalle et al., 2020)

- 🔒 Total number of problems ( $m$ )?
- 🔒 Probabilities of each problem ( $p_1, \dots, p_m$ )?
- 🔒 Economic consequences of undiscovered problems?
- 🔒 Number of new patients ( $n'$ ) needed to limit economic consequences?
- 🔒 Model the discovery matrix
- 🔒 Heterogeneous discovery probabilities
- 🔒 Fully Bayesian: MCMC
- 🔒 Include in value of information framework to assess the number of new patients needed

$$n \text{ patients} \left\{ \begin{array}{c} \overbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}}^{j \text{ discovered problems}} \end{array} \right.$$



Update performance tables based on flight data (Dewez, Guedj, and Vandewalle, 2020)

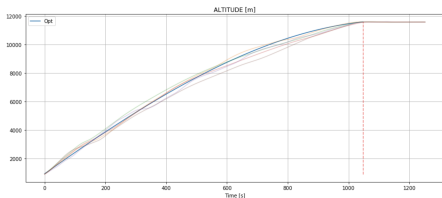
- 🔒 Unobserved forces
- 🔒 Physical approximations
- 🔒 Relevance of the approximation
- 🔒 Bounds on the error

Optimizing through a functional approach

- 🔒 End-to-end optimization
- 🔒 Functional approach
- 🔒 Include constraints
- 🔒 Penalty based on reference trajectories



Source: [traveller.com.au](http://traveller.com.au)

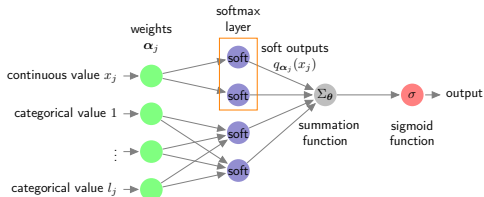
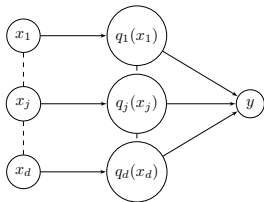


## Rational review of reject inference methods

- How to use data of non-financed applicants?
- Empirical procedures
- Re-interpretation in missing data framework
- Analysis of reject inference methods

## Model-embedded feature quantization

- Prior feature quantization
- Combinatorial challenge
- Embed feature quantization
- Continuous relaxation
- Neural network-based estimation





1. Introduction
2. Contribution to model-based clustering
  - General presentation
  - Focus on multiple partition clustering
3. Contribution to some applications
4. Research project

## Multiple partition clustering dissemination

- 🔒 Number of blocks and number of clusters in each block
- 🔧 Disseminate a user-friendly toolbox

## Predictive clustering

- 🔒 Embed clustering step in predictive models
- 🔒 Consider multiple partition in predictive clustering

## Clustering of recurrent event data and integration of covariates

- 🔒 Cluster recurrent event data: application to re-hospitalization
- 🔒 Include covariates in the clustering model  $\Rightarrow$  added value clustering

## Application to medical data

- 🔒 Temporal structure in analysis of high-throughput proteomic data (Wilfried Heyse thesis)
- 🔒 Model patient path at hospital







## Application to industry/retail







- 🔒 Sales prediction by grouping low turnover product (Axel Potier thesis)
- 🔒 Data-driven trajectory optimization






*Classification has consequences.*

— David Hand, IFCS 2019

- Obvious in supervised classification, but also true in clustering
- Global issue but only local solutions
- 🔒 Include information from the problem when performing clustering
- 👤 Focus on collaborative projects in medicine
- 🔒 Encompass clustering in a dynamic decision making process

-  Attias, Hagai (1999). “Independent factor analysis”. In: *Neural computation* 11.4, pp. 803–851.
-  Biernacki, Christophe, Matthieu Marbac, and Vincent Vandewalle (2020). “Gaussian-Based Visualization of Gaussian and Non-Gaussian-Based Clustering”. In: *Journal of Classification*.
-  Biernacki, Christophe and Vincent Vandewalle (2011). “Label switching in mixtures”. In: *AIP Conference Proceedings*. Vol. 1389. 1. American Institute of Physics, pp. 398–401.
-  Dewez, Florent, Benjamin Guedj, and Vincent Vandewalle (2020). “From industry-wide parameters to aircraft-centric on-flight inference: Improving aeronautics performance prediction with machine learning”. In: *Data-Centric Engineering* 1, e11.
-  Dhaenens, Clarisse et al. (2018). “ClinMine: Optimizing the management of patients in hospital”. In: *IRBM* 39.2, pp. 83–92.
-  Eirola, Emil et al. (2014). “Mixture of Gaussians for distance estimation with missing data”. In: *Neurocomputing* 131, pp. 32–42.

-  Galimberti, Giuliano, Annamaria Manisi, and Gabriele Soffritti (2018). “Modelling the role of variables in model-based cluster analysis”. In: *Statistics and Computing* 28.1, pp. 145–169.
-  Galimberti, Giuliano and Gabriele Soffritti (2007). “Model-based methods to identify multiple cluster structures in a data set”. In: *Computational Statistics & Data Analysis* 52.1, pp. 520 –536.
-  Marbac, Matthieu, Christophe Biernacki, and Vincent Vandewalle (2015). “Model-based clustering for conditionally correlated categorical data”. In: *Journal of Classification* 32.2, pp. 145–175.
-  – (2016). “Latent class model with conditional dependency per modes to cluster categorical data”. In: *Advances in Data Analysis and Classification* 10.2, pp. 183–207.
-  – (2017). “Model-based clustering of Gaussian copulas for mixed data”. In: *Communications in Statistics - Theory and Methods* 46.23, pp. 11635–11656.
-  Marbac, Matthieu and Mohammed Sedki (2017). “Variable selection for model-based clustering using the integrated complete-data likelihood”. In: *Statistics and Computing* 27.4, pp. 1049–1063.

-  Marbac, Matthieu and Vincent Vandewalle (2019). “A tractable multi-partitions clustering”. In: *Computational Statistics & Data Analysis* 132, pp. 167–179.
-  Poon, Leonard K. M. et al. (2013). “Model-based clustering of high-dimensional data: Variable selection versus facet determination”. In: *International Journal of Approximate Reasoning* 54.1, pp. 196–215.
-  Vandewalle, Vincent (2020). “Multi-Partitions Subspace Clustering”. In: *Mathematics* 8.4, p. 597.
-  Vandewalle, Vincent and Christophe Biernacki (Dec. 2015). “An efficient SEM algorithm for Gaussian Mixtures with missing data”. In: *8th International Conference of the ERCIM WG on Computational and Methodological Statistics*. Londres, United Kingdom.
-  Vandewalle, Vincent, Cristian Preda, and Sophie Dabo (2020). “Clustering spatial functional data”. In: *Geostatistical Functional Data Analysis : Theory and Methods*. Ed. by J. Mateu and R. Giraldo. ISBN: 978-1-119-38784-8. Chichester, UK: John Wiley and Sons.



Vandewalle, Vincent et al. (2020). “Estimating the number of usability problems affecting medical devices: modelling the discovery matrix”. In: *BMC Medical Research Methodology* 20.234.



