Model-based clustering of categorical data by relaxing conditional independence

M. Marbac[3,6], C. Biernacki[3,4,5], V. Vandewalle[1,2,3]

Classification society meeting 2015
Mc Master University

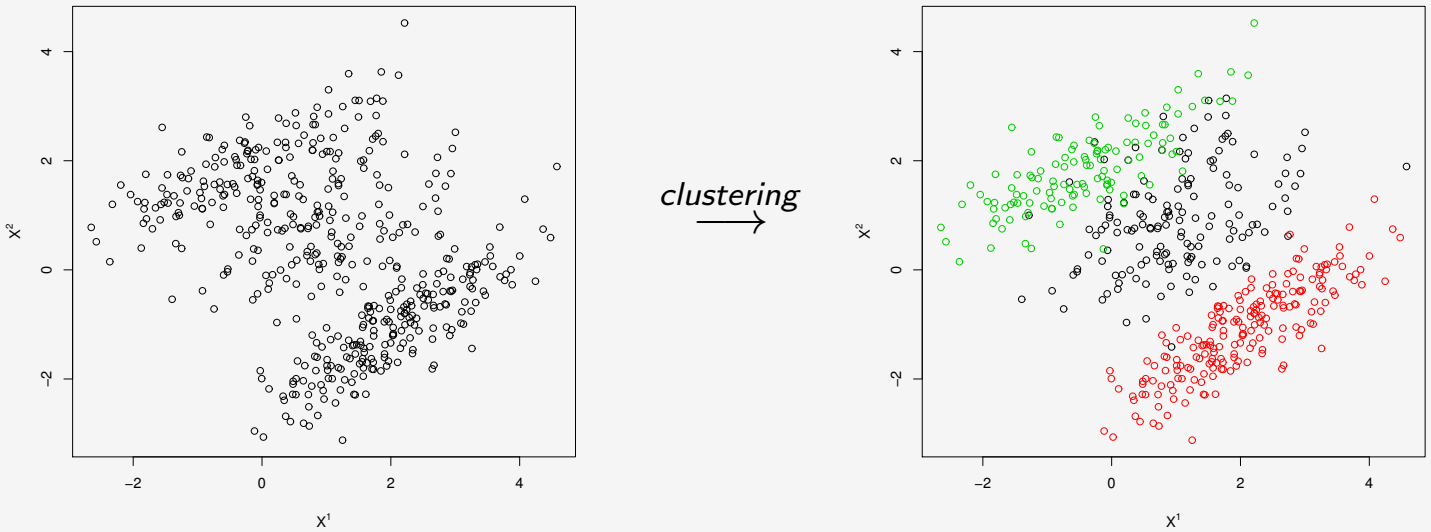
5 June 2015

1 Université de Lille 2 DROIT ET SANTÉ
2 équipe d'accueil EA 2694
3 Inria informatics mathematics
4 cnrs dépasser les frontières
5 Université de Lille 1 SCIENCES ET TECHNOLOGIES
6 DGA

Outline

Model-based clustering

$\mathbf{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ $\hat{\mathbf{z}} = (\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n)$, $\hat{g}$ clusters



Mixture model: well-posed problem

$$p(\mathbf{x}; \boldsymbol{\theta}|g) = \sum_{k=1}^{g} \pi_k p(\mathbf{x}; \boldsymbol{\theta}_k|g) \quad \text{can be used for} \quad \begin{cases} \mathbf{x} \to \hat{\boldsymbol{\theta}} \to p(\mathbf{z}|\mathbf{x}, g; \hat{\boldsymbol{\theta}}) \to \hat{\mathbf{z}} \\ \mathbf{x} \to \hat{p}(g|\mathbf{x}) \to \hat{g} \end{cases}$$

with $\boldsymbol{\theta} = ((\pi_1, \ldots, \pi_k, \ldots, \pi_g), (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k, \ldots, \boldsymbol{\alpha}_g))$

Categorical data

d categorical variables, each with m_j response levels

- $\boldsymbol{x}_i = \{x_i^j : j = 1, \ldots, d\}$
- $x_i^j = \{x_i^{jh} : h = 1, \ldots, m_j\}$
- $x_i^{jh} = 1$ if i has response level h for variable j and $x_i^{jh} = 0$ otherwise

Example ("Genes Diffusion" company):

- $n = 4270$ calves
- $d = 9$ variables of behavior[1] and health related[2]
- Response levels of TRC ($j = 3$): TRC$\in\{$"curative", "preventive", "no"$\}$ ($m_3 = 3$)

$$\begin{array}{lcll} \boldsymbol{x}_1^3 & = & \text{"curative"} & = (1\ 0\ 0) \\ \boldsymbol{x}_2^3 & = & \text{"no"} & = (0\ 0\ 1) \\ \boldsymbol{x}_3^3 & = & \text{"no"} & = (0\ 0\ 1) \\ \vdots & \vdots & \vdots & \vdots \quad \vdots \end{array}$$

[1] aptitude for sucking *Apt*, behavior of the mother just before the calving *Iso*

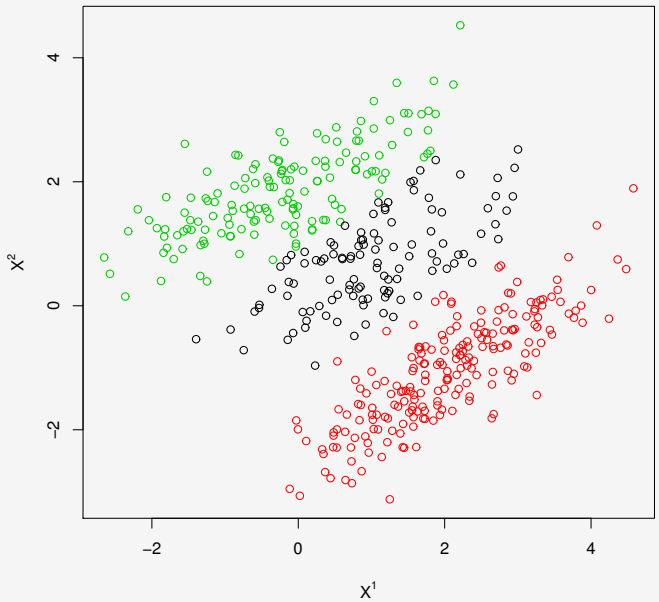
[2] treatment against omphalite *TOC*, respiratory disease *TRC* and diarrhea *TDC*, umbilicus disinfection *Dis*, umbilicus emptying *Emp*, mother preventive treatment against respiratory disease *TRM* and diarrhea *TDM*

Intra-class correlations

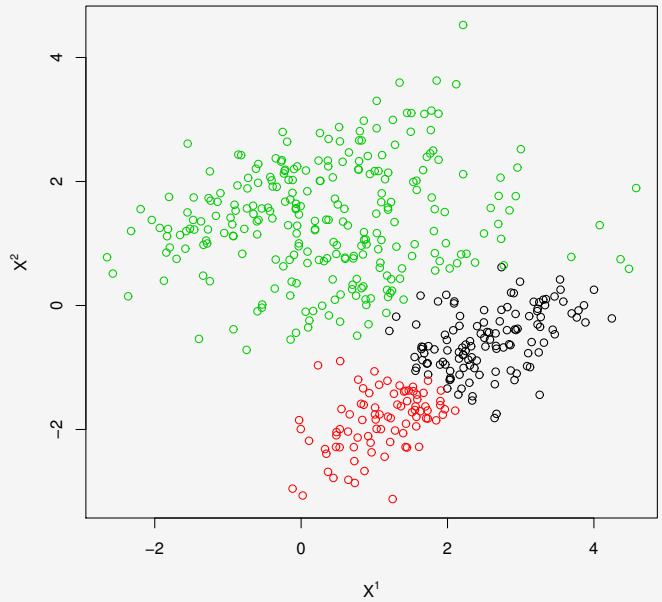
A nowadays interest

- More frequent (in the population) when d increases
- More observable (in the sample) when n increases
- Risk of bias when models do not take into account such correlations

Bias example (on $\mathbf{z}$) with Gaussians:



Correlated Gaussians



Independent Gaussians

Classical categorical models

- Conditional independence (CIM): linked to some χ^2 distance-based methods

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

where $\boldsymbol{\alpha}_k = \{\alpha_k^{jh} : j = 1,\ldots,d, h = 1,\ldots,m_j\}$ and $\alpha_k^{jh} = p(x^{jh} = 1|\boldsymbol{z} = k)$

$\ominus$ bias

- Dependence trees: allows only certain dependencies

$\ominus$ too many parameters and unstable estimation of the tree

- Latent Trait Analyzers: a continuous variable explains intra-dependency

$$p(\boldsymbol{x};\boldsymbol{\alpha}_k) = \int_{\mathbb{R}^{|\boldsymbol{c}|}} \prod_{j=1}^{d}\prod_{h=1}^{m_j} p(x^{jh}|\boldsymbol{c};\boldsymbol{\alpha}_k)p(\boldsymbol{c})d\boldsymbol{c}$$

$\ominus$ difficult to meaningfully explain correlations

The "gold rule"

A model should be flexible + parsimonious + meaningful

Dependence per blocks (1/3)

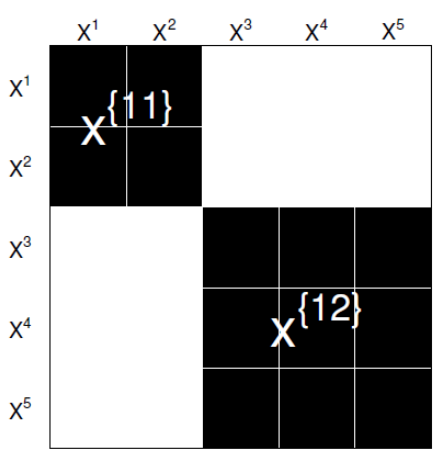
- *Conditionally* on the class k, variables are grouped into B_k *independent blocks*
- Partition of variables: $\boldsymbol{\sigma}_k = (\boldsymbol{\sigma}_{k1}, \ldots, \boldsymbol{\sigma}_{k\mathrm{B}_k})$ of $\{1, \ldots, d\}$
- Number of variables in the block b of the component k: $d^{\{kb\}} = \mathrm{card}(\boldsymbol{\sigma}_{kb})$
- Subset of $\boldsymbol{x}$ associated to $\boldsymbol{\sigma}_{kb}$: $\boldsymbol{x}^{\{kb\}} = \boldsymbol{x}^{\boldsymbol{\sigma}_{kb}} = (\boldsymbol{x}^{\{kb\}j}; j = 1, \ldots, d^{\{kb\}})$
- Variable j of the block b for component k: $\boldsymbol{x}^{\{kb\}j} = (x^{\{kb\}jh}; h = 1, \ldots, m_j^{\{kb\}})$
- Modalities number of $\boldsymbol{x}^{\{kb\}j}$: $m_j^{\{kb\}}$
- All repartitions in blocks: $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_g)$
- Distribution per class:

$$p(\boldsymbol{x}; \boldsymbol{\theta}_k | \boldsymbol{\sigma}_k, g) = \prod_{b=1}^{B_k} p(\boldsymbol{x}^{\{kb\}}; \boldsymbol{\theta}_{kb}) \quad \text{with} \quad \boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{k\mathrm{B}_k})$$

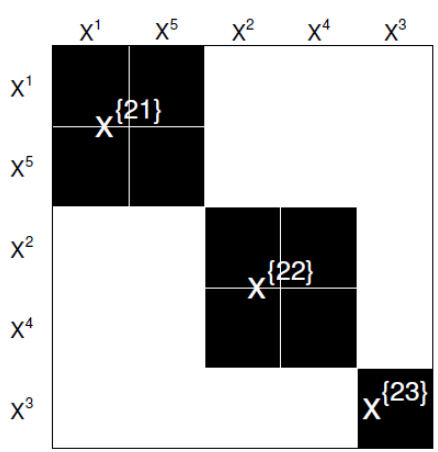
Inter-Block model $\boldsymbol{\sigma}_k$ verifies the "gold rule"

Dependence per blocks (2/3)

Example with $g = 2$, $d = 5$:



$k = 1$, $B_1 = 2$
$\boldsymbol{\sigma}_1 = (\{1, 2\}, \{3, 4, 5\})$



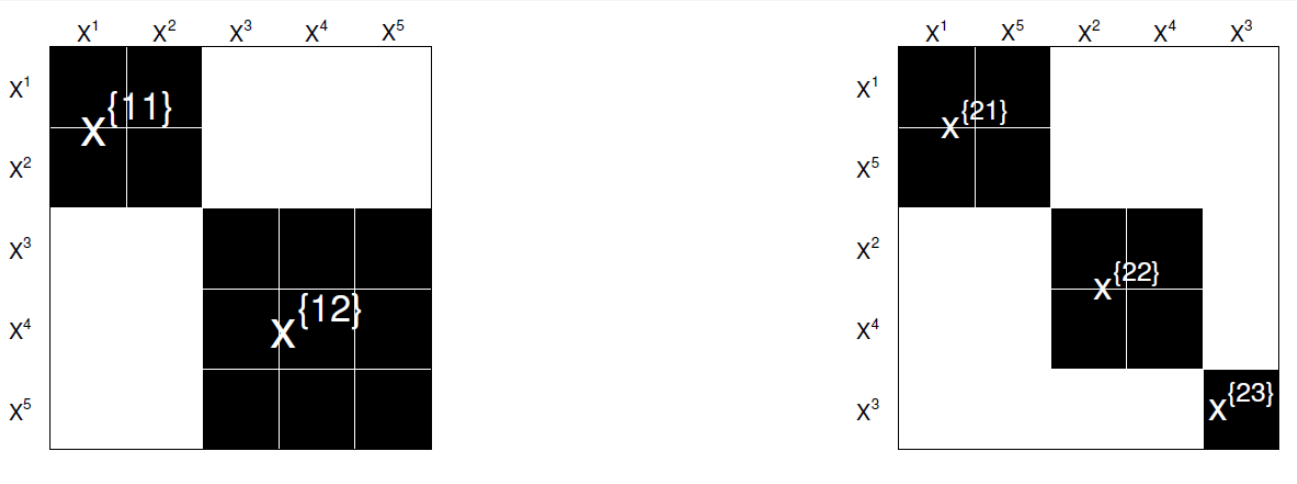
$k = 2$, $B_2 = 3$
$\boldsymbol{\sigma}_2 = (\{1, 5\}, \{2, 4\}, \{3\})$

The present work

Intra-block distribution $p(\boldsymbol{x}^{\{kb\}}; \boldsymbol{\theta}_{kb})$ should also verify the "gold rule"

Dependence per blocks (2/3)

Example with $g = 2$, $d = 5$:



$k = 1$, $B_1 = 2$
$\boldsymbol{\sigma}_1 = (\{1, 2\}, \{3, 4, 5\})$

$k = 2$, $B_2 = 3$
$\boldsymbol{\sigma}_2 = (\{1, 5\}, \{2, 4\}, \{3\})$

The present work

Two Intra-block distributions are now proposed...

Outline

Maximum dependency distribution

Main idea

- The "opposite" distribution of independence according to the Cramer's V criterion computed on all the couples of variables
- The knowledge of the variable having the largest number of modalities determines exactly the others

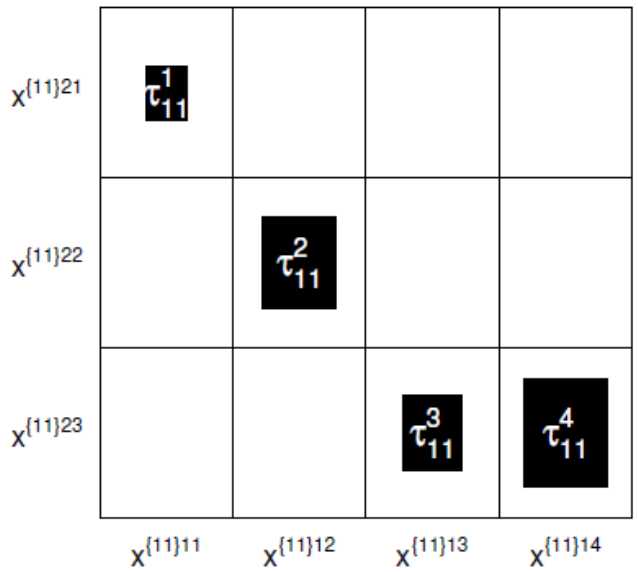
- Variables are ordered by decreasing number of modalities in each block
- Successive surjections from the space of $x^{\{kb\}j}$ to the space of $x^{\{kb\}j+1}$

$$p(\boldsymbol{x}^{\{kb\}};\boldsymbol{\tau}_{kb},\boldsymbol{\delta}_{kb}) = \overbrace{p(\boldsymbol{x}^{\{kb\}1};\boldsymbol{\tau}_{kb})}^{\text{1st variable}} \overbrace{\prod_{j=2}^{d^{\{kb\}}} p(\boldsymbol{x}^{\{kb\}j}|\boldsymbol{x}^{\{kb\}1};\{\boldsymbol{\delta}_{kb}^{hj}\}_{h=1,\dots,m_1^{\{kb\}}})}^{\text{other variables}}$$

$$= \prod_{h=1}^{m_1^{\{kb\}}} \left(\underbrace{\tau_{kb}^{h}}_{\in(0,1)} \prod_{j=2}^{d^{\{kb\}}} \prod_{h'=1}^{m_j^{\{kb\}}} \left(\underbrace{\delta_{kb}^{hjh'}}_{\in\{0,1\}} \right)^{x^{\{kb\}jh'}} \right)^{x^{\{kb\}1h}}$$

with $\boldsymbol{\delta}_{kb} = (\boldsymbol{\delta}_{kb}^{hj})$, $\boldsymbol{\delta}_{kb}^{hj} = (\delta_{kb}^{hjh'})$, $\boldsymbol{\tau}_{kb} = (\tau_{kb}^{h})$

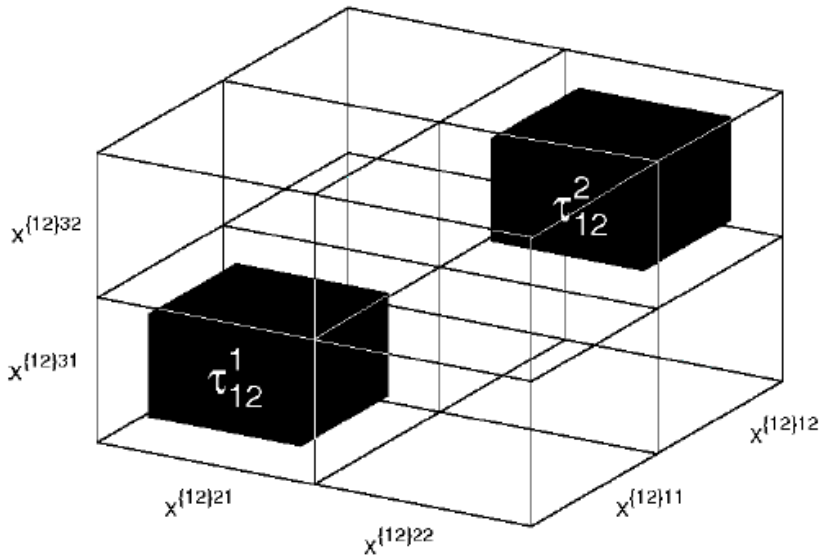
Example



$m^{\{11\}1} = 4,\ m^{\{11\}2} = 3$

$\delta^{h1h}_{11} = 1$ for $h = 1, 2, 3,\ \delta^{413}_{11} = 1$

$\boldsymbol{\tau}_{11} = (0.1, 0.3, 0.2, 0.4)$



$m^{\{12\}1} = m^{\{12\}2} = m^{\{12\}3} = 2$

$\delta^{hjh'}_{12} = 1$ iff $(h = h')$

$\boldsymbol{\tau}_{12} = (0.5, 0.5)$

Mixture of extreme distributions (CCM1)

CCM1

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

where $\boldsymbol{\theta}_{kb} = (\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$

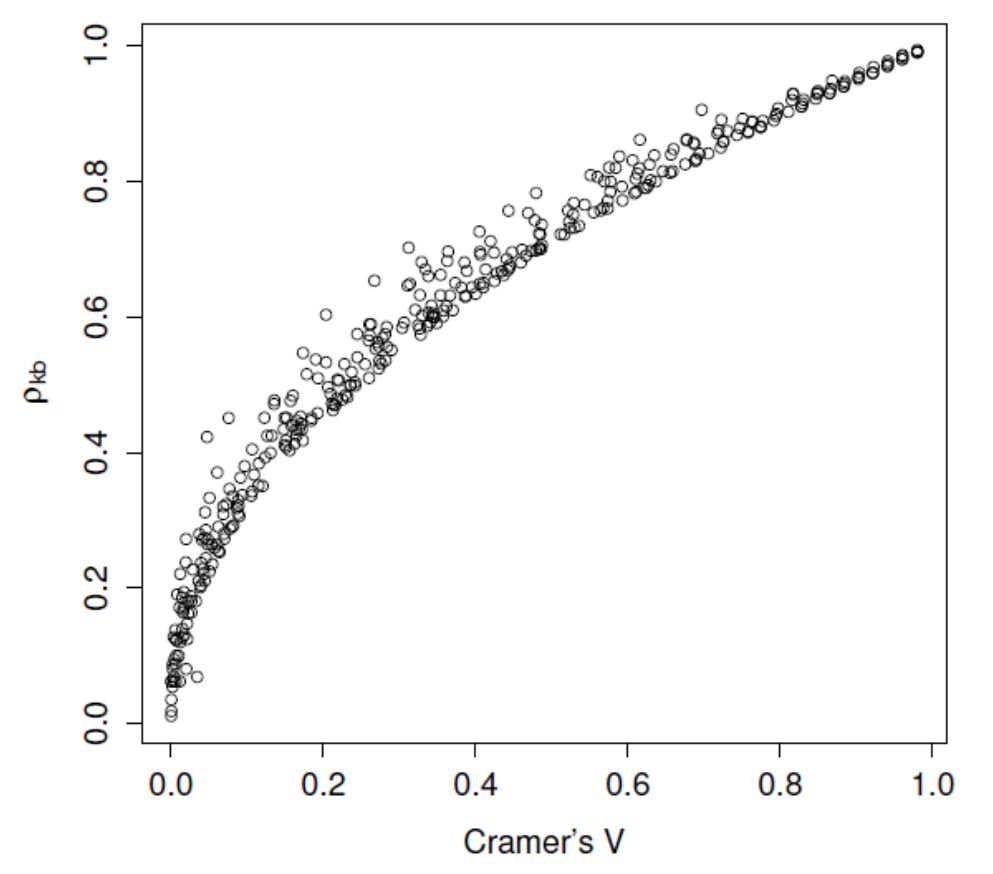
- Meaningful:
  - ρ_{kb}: global inter-variable correlation in the block ($0 \leq \rho_{kb} \leq 1$)
  - $\boldsymbol{\delta}_{kb}$: intra-variable correlation in the block ($\in \{0,1\}$)
- Parsimony:

$$\nu_{\text{CCM1}} = \nu_{\text{CIM}} + \underbrace{\sum_{\{(k,b)|d^{\{kb\}}>1\}} m_1^{\{kb\}}}_{\text{nb modalities of the 1st variable in the block}}$$

- Identifiable if $d^{\{kb\}} > 2$ or $m_2^{\{kb\}} > 2$ (additional constraints added otherwise)

ρ_{kb} *vs.* Cramer's V

Empirical link between ρ_{kb} and the Cramer's V for two binary variables



Estimation of $\boldsymbol{\theta}$ (1/3)

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}|g, \boldsymbol{\sigma}) \quad \text{with model } (g, \boldsymbol{\sigma}) \text{ fixed}$$

Global GEM algorithm

- **$\mathbf{E}_{\text{global}}$ step:**

$$z_{ik}^{(r)} = \frac{\pi_k^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^{g} \pi_{k'}^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\sigma}_{k'}, \boldsymbol{\theta}_{k'}^{(r)})}$$

- **$\mathbf{GM}_{\text{global}}$ step:**

$$\pi_k^{(r+1)} = \frac{n_k^{(r)}}{n} \quad \text{where} \quad n_k^{(r)} = \sum_{i=1}^{n} z_{ik}^{(r)}$$

$$\forall (k, b), \quad \boldsymbol{\theta}_{kb}^{(r+1)} = \text{argmax}_{\boldsymbol{\theta}_{kb}} L(\boldsymbol{\theta}_{kb}; \mathbf{x}, \mathbf{z}^{(r)}|g, \boldsymbol{\sigma}) \quad \longrightarrow \text{MH algorithm}$$

Estimation of $\boldsymbol{\theta}$ (2/3)

$$\forall (k,b), \quad \boldsymbol{\theta}_{kb}^{(r+1)} = \text{argmax}_{\boldsymbol{\theta}_{kb}} L(\boldsymbol{\theta}_{kb}; \mathbf{x}, \mathbf{z}^{(r)} | g, \boldsymbol{\sigma}) \quad \text{with } (\mathbf{z}^{(r)}, g, \boldsymbol{\sigma}) \text{ fixed}$$

Metropolis-Hastings algorithm (discrete parameters $\boldsymbol{\delta}_{kb}$)

- **Proposal distribution:**

$$\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})} \sim \text{uniform distribution in a neighborhood } \Delta(\boldsymbol{\delta}_{kb}^{(r,s)})$$

$$(\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb})^{(r,s+\frac{1}{2})} = \text{argmax}_{\bullet} L(\bullet; \mathbf{x}, \mathbf{z}^{(r)}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})} | g, \boldsymbol{\sigma}) \quad \longrightarrow \text{EM algorithm}$$

- **Acceptance distribution:**

$$\mu^{(r,s+1)} = \min \left\{ \frac{\prod_{i=1}^{n} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})})^{z_{ik}^{(r)}} |\Delta(\boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})|}{\prod_{i=1}^{n} p(\boldsymbol{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}^{(r,s)})^{z_{ik}^{(r)}} |\Delta(\boldsymbol{\delta}_{kb}^{(r,s)})|}, 1 \right\}$$

$$\boldsymbol{\theta}_{kb}^{(r,s+1)} = \begin{cases} \boldsymbol{\theta}_{kb}^{(r,s+\frac{1}{2})} & \text{with probability } \mu^{(r,s+1)} \\ \boldsymbol{\theta}_{kb}^{(r,s)} & \text{otherwise} \end{cases}$$

Estimation of $\boldsymbol{\theta}$ (3/3)

$(\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb})^{(r,s+\frac{1}{2})} = \text{argmax}_{\bullet} L(\bullet; \mathbf{x}, \mathbf{z}^{(r)}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})} | g, \boldsymbol{\sigma})$ with $(\mathbf{z}^{(r)}, g, \boldsymbol{\sigma}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})$ fixed

New latent variable: (with Bernoulli distribution)

- $y_i^{\{kb\}} = 1$: $\mathbf{x}_i^{\{kb\}} \sim$ *maximum dependency* distribution for block b of cluster k
- $y_i^{\{kb\}} = 0$: $\mathbf{x}_i^{\{kb\}} \sim$ *independence* distribution for block b of cluster k
- $\mathbf{y} = (\mathbf{y}^{\{kb\}}; k = 1, \ldots, g; b = 1, \ldots, \mathrm{B}_k)$ with $\mathbf{y}^{\{kb\}} = (y_1^{\{kb\}}, ..., y_n^{\{kb\}})$

EM algorithm (mixture independence / extreme dependency)

- **$\mathbf{E}_{\text{local}}$ step:** $y_i^{\{kb\}(r,s+\frac{1}{2},t)} \propto \rho_{kb}^{(r,s+\frac{1}{2},t)} p(\mathbf{x}_i^{\{kb\}}; \boldsymbol{\tau}_{kb}^{(r,s+\frac{1}{2},t)}, \boldsymbol{\delta}_{kb}^{(r,s+\frac{1}{2})})$
- **$\mathbf{M}_{\text{local}}$ step:** $\rho_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{n_{kb}^{(r,s+\frac{1}{2},t)}}{n_k^{(r)}}$, $\tau_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{\sum_{i=1}^n z_{ik}^{(r)} y_i^{\{kb\}(r,s+\frac{1}{2},t)} x_i^{\{kb\}1h}}{n_{kb}^{(r,s+\frac{1}{2},t)}}$,

$\alpha_{kb}^{(r,s+\frac{1}{2},t+1)} = \frac{\sum_{i=1}^n z_{ik}^{(r)} (1 - y_i^{\{kb\}(r,s+\frac{1}{2},t)}) x_i^{\{kb\}jh}}{n_k^{(r)} - n_{kb}^{(r,s+\frac{1}{2},t)}}$, where $n_{kb}^{(r,s+\frac{1}{2},t)} = \sum_{i=1}^n z_{ik}^{(r)} y_i^{\{kb\}(r,s+\frac{1}{2},t)}$

Conjecture: Unique maximum!

Model selection

$$(\hat{g}, \hat{\boldsymbol{\sigma}}) = \underset{g,\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(g, \boldsymbol{\sigma}|\mathbf{x}) = \underset{g}{\operatorname{argmax}} \left[\underset{\boldsymbol{\sigma}}{\operatorname{argmax}}\ p(\mathbf{x}|g, \boldsymbol{\sigma})\right]$$

Gibbs algorithm (as a reversible jump)

- **Neighborhood step:**
$$\Sigma^{[q]} \sim \Sigma|\boldsymbol{\sigma}^{[q]}$$
- **Pattern step:**
$$\boldsymbol{\sigma}^{[q+1]} \sim p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]})$$

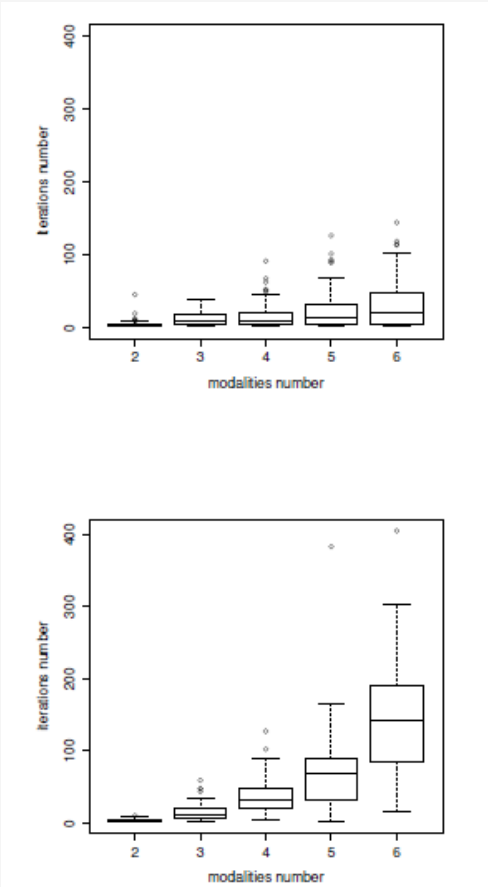
with
$$p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]}) = \begin{cases} \frac{p(\mathbf{x}|g,\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}' \in \Sigma^{[q]}} p(\mathbf{x}|g,\boldsymbol{\sigma}')} & \text{if } \boldsymbol{\sigma} \in \Sigma^{[q]} \\ 0 & \text{otherwise.} \end{cases}$$

and using the BIC approximation
$$\ln p(\mathbf{x}|g, \boldsymbol{\sigma}) \simeq L(\hat{\boldsymbol{\theta}}; \mathbf{x}|g, \boldsymbol{\sigma}) - \frac{\nu_{\text{CCM1}}}{2} \log(n),$$

Tuning

- Initialization: HAC on the matrix of Cramer's distances on the couples of variables
- Stopping criteria:

Algorithms	Gibbs	GEM	Metropolis-Hastings	EM
Criteria	$q_{\max} = 20 \times d$	$r_{\max} = 10$	$s_{\max} = 1$	$t_{\max} = 5$



$\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}$ with/without init.

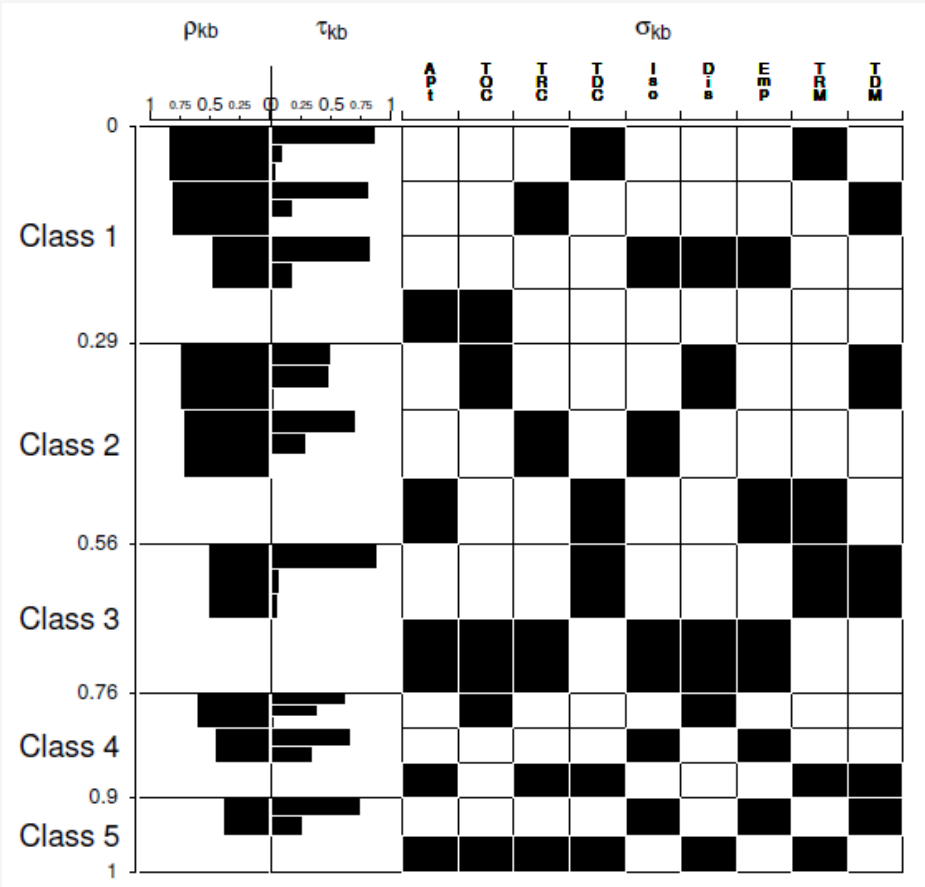
$d \setminus n$	100	200	400	800
4	**0.77** (*1.34*)	**0.26** (*0.26*)	**0.15** (*0.05*)	**0.12** (*0.05*)
6	**1.22** (*1.77*)	**0.27** (*0.14*)	**0.09** (*0.07*)	**0.05** (*0.05*)
8	**1.72** (*2.50*)	**0.41** (*0.20*)	**0.09** (*0.05*)	**0.05** (*0.03*)
10	**1.73** (*4.06*)	**0.52** (*0.14*)	**0.10** (*0.03*)	**0.04** (*0.03*)

mean (*standard deviation*) of the Kullback-Leibler divergence.

Model selection

Calves (1/2)

	g	1	2	3	4	5	6	7	8
CIM	BIC	-28589	-26859	-26526	-26333	-26238	-26235	-26226	**-26185**
	ν_{CIM}	17	35	53	71	89	107	125	**143**
CCM1	BIC	-26653	-26289	-26173	-26038	**-26025**	-26059	-26045	-26058
	ν_{CCM1}	24	48	80	89	**112**	131	148	163
	time (min)	0.97	3.32	6.16	6.56	**10.03**	11.76	12.31	14.92



Calves (2/2)

The first class has a proportion of 0.29 and it is composed of four blocks. The most correlated block of the first class has $\rho_{kb} \simeq 0.80$ and the strength of the biggest modalities link is close to 0.85 too. This block consists of the variables *TDC* and *TRM*. Here is now a possible interpretation of Class 1:

- **General:** this class has a proportion equal to 0.29 and consists of three blocks of dependency and one block of independence.
- **Block 1:** there is a strong correlation (ρ_{11}) between the variables diarrhea treatment of the calve and mother preventive treatment against respiratory disease, especially between the modality no treatment against the calve diarrhea and the absence of preventive treatment against respiratory disease of its mother ($\boldsymbol{\tau}_{11}$ and $\boldsymbol{\delta}_{11}$).
- **Block 2:** there is a strong correlation (ρ_{12}) between the variables treatment against respiratory illness of the calve and mother preventive treatment against diarrhea, especially between the modality preventive treatment against respiratory illness of the calve and the presence of diarrhea preventive treatment of its mother ($\boldsymbol{\tau}_{12}$ and $\boldsymbol{\delta}_{12}$).
- **Block 3:** there exists another strong link between the behavior of the mother, the emptying of the umbilical and its disinfection ($\boldsymbol{\tau}_{13}$ and $\boldsymbol{\delta}_{13}$).
- **Block 4:** this block is characterized by absence of preventive treatment against omphalite and have 50% of the calves infected by this illness ($\boldsymbol{\alpha}_{14}$).

Dentistry (1/2)

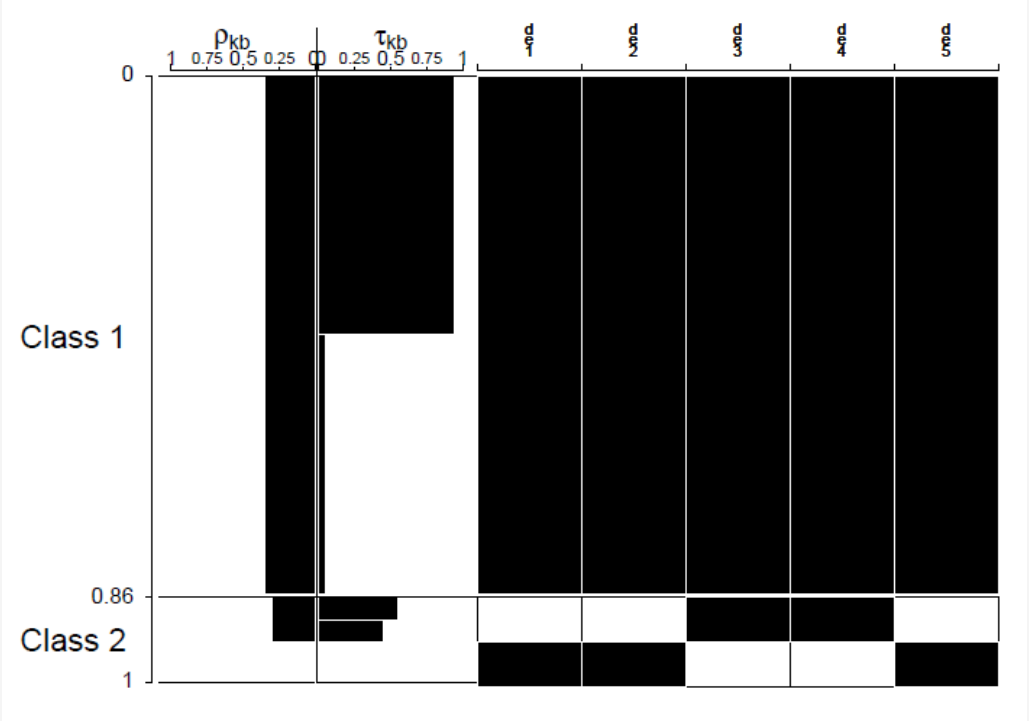
- $n = 3869$ dental x-rays (sound or carious) evaluated by $d = 5$ dentists
- Past experiments suggested two main classes: sound teeth and carious ones

Dentist						Dentist					
1	2	3	4	5	Frequency	1	2	3	4	5	Frequency
S	S	S	S	S	1880	C	S	S	S	S	22
S	S	S	S	C	789	C	S	S	S	C	26
S	S	S	C	S	43	C	S	S	C	S	6
S	S	S	C	C	75	C	S	S	C	C	14
S	S	C	S	S	23	C	S	C	S	S	1
S	S	C	S	C	63	C	S	C	S	C	20
S	S	C	C	S	8	C	S	C	C	S	2
S	S	C	C	C	22	C	S	C	C	C	17
S	C	S	S	S	188	C	C	S	S	S	2
S	C	S	S	C	191	C	C	S	S	C	20
S	C	S	C	S	17	C	C	S	C	S	6
S	C	S	C	C	67	C	C	S	C	C	27
S	C	C	S	S	15	C	C	C	S	S	3
S	C	C	S	C	85	C	C	C	S	C	72
S	C	C	C	S	8	C	C	C	C	S	1
S	C	C	C	C	56	C	C	C	C	C	100

Radiographic cross-diagnosis of 3869 molars and premolars by five dentists

	g	1	2	3	4
CIM	BIC	-8766	-7511	**-7481**	-7503
CCM1	BIC	-7743	**-7473**	-7481	-7503
	time (sec)	1.7	4.9	6.1	7.7

Dentistry (2/2)



- the majority class ($\pi_1 = 0.86$) mainly gathers the sound teeth. There is a strong dependency between the five dentists ($\boldsymbol{\sigma}_1 = (\{1, 2, 3, 4, 5\})$ and $\rho_{11} = 0.35$). The dependency structure of the maximum dependency distribution indicates an over contribution of both modality interactions where the five dentists have the same diagnosis, especially when they claim that the teeth is sound ($\tau_{11}^{\text{all\_sound}} = 0.93$ and $\tau_{11}^{\text{all\_carious}} = 0.07$).
- the minority class ($\pi_2 = 0.14$) groups principally the carious teeth. There is a dependency between the dentists 3 and 4 while the diagnosis of the other ones are independent given the class ($\boldsymbol{\sigma}_2 = (\{3, 4\}, \{1, 2, 5\})$, $\rho_{21} = 0.31$ and $\rho_{22} = 0$).

Outline

Dependence per blocks

Restriction (for identifiability)

Blocks are equal per class

- *Conditionally* on all classes, variables are grouped into B *independent blocks*
- Partition of variables: $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{\mathrm{B}})$ of $\{1, \ldots, d\}$
- Number of variables in the block b of all components: $d^{\{b\}} = \mathrm{card}(\boldsymbol{\sigma}_b)$
- Subset of $\boldsymbol{x}$ associated to $\boldsymbol{\sigma}_b$: $\boldsymbol{x}^{\{b\}} = \boldsymbol{x}^{\boldsymbol{\sigma}_b} = (\boldsymbol{x}^{\{b\}j}; j = 1, \ldots, d^{\{b\}})$
- Variable j of the block b for all components: $\boldsymbol{x}^{\{b\}} = (x^{\{b\}h}; h = 1, \ldots, m^{\{b\}})$
- Modalities number of $\boldsymbol{x}^{\{b\}}$: $m^{\{b\}} = \prod_{j=1}^{d^{\{b\}}} m_j^{\{b\}}$
- Distribution per class:

$$p(\boldsymbol{x}; \boldsymbol{\alpha}_k | \boldsymbol{\sigma}, g) = \prod_{b=1}^{B} p(\boldsymbol{x}^{\{b\}}; \boldsymbol{\alpha}_{kb}) \quad \text{with} \quad \boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1}, \ldots, \boldsymbol{\alpha}_{k\mathrm{B}})$$

with $\boldsymbol{\alpha}_{kb} = (\alpha_{kb}^h; h = 1, \ldots, m^{\{b\}})$

Conditional dependency modes distribution (CCM2)

Main idea

- The distribution of modalitiy crossings in each block is uniform
- Except modes: some particular modality crossings with higher (free) probability

- Number of modes in block b, class k: ℓ_{kb}
- Number of modes in class k: $\boldsymbol{\ell}_k = (\ell_{k1}, \ldots, \ell_{k\mathrm{B}})$
- All number of modes: $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_g)$
- The model:

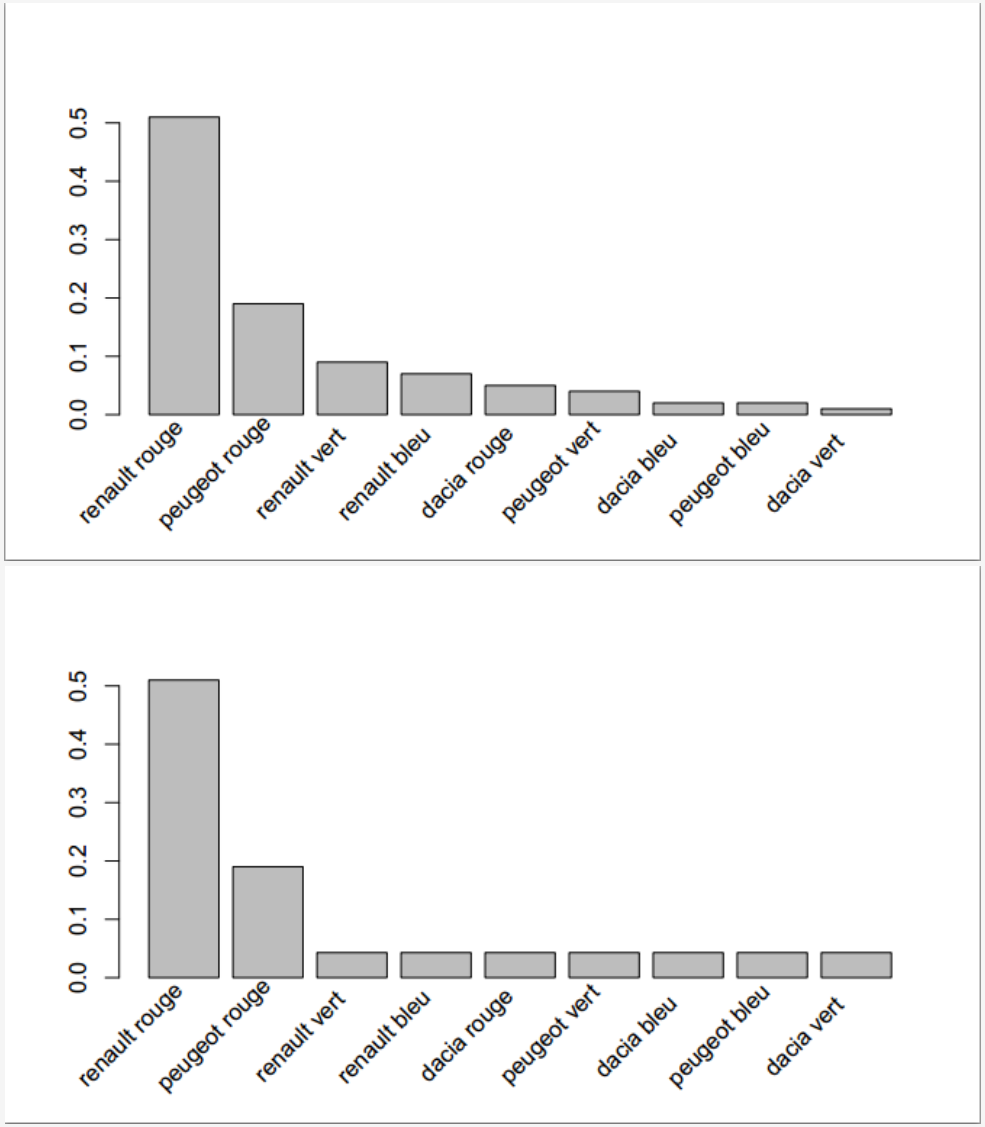
$$p(\boldsymbol{x}_i^{\{b\}}; \boldsymbol{\alpha}_{kb}, \ell_{kb}) = \prod_{h=1}^{m^{\{b\}}} \left(\alpha_{kb}^h\right)^{x_i^{\{b\}h}}$$

with

$$0 \leq \alpha_{kb}^h \leq 1, \ \sum_{h=1}^{m^{\{b\}}} \alpha_{kb}^h = 1, \ \alpha_{kb}^{(\ell_{kj}+1)} = \ldots = \alpha_{kb}^{(m^{\{b\}})}$$

where the elements of $\boldsymbol{\alpha}_{kb}$ are ordered by decreasing values: $\alpha_{kb}^{(h)} \geq \alpha_{kb}^{(h+1)}$

Illustration of a CCM2 block



Properties of CCM2

- Identifiability
- Parsimony:

$$\nu_{\mathrm{CMM2}} = (g-1) + \sum_{k=1}^{g}\sum_{b=1}^{\mathrm{B}} \ell_{kb} \leq (g-1) + g \times \sum_{b=1}^{\mathrm{B}} (m^{\{b\}} - 1)$$

- Meaningful:
  - Intra-variable dependencies described by modes (locations and probabilities)
  - Complexity of intra-variable dependencies:

$$\kappa_{kb} = \frac{\ell_{kb}}{m^{\{b\}} - 1}$$

  - Stength of intra-variable dependencies:

$$\tau_{kb} = \sum_{h=1}^{\ell_{kb}} \alpha_{kb}^{(h)}$$

Estimation of $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x} | g, \boldsymbol{\sigma}, \boldsymbol{\ell}) \quad \text{with model } (g, \boldsymbol{\sigma}, \boldsymbol{\ell}) \text{ fixed}$$

EM algorithm

- **E step:** conditional probabilities computation

$$t_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{[r]}, \boldsymbol{\sigma}, \boldsymbol{\ell}_k)}{\sum_{k'=1}^{g} \pi_{k'}^{[r]} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_{k'}^{[r]}, \boldsymbol{\sigma}, \boldsymbol{\ell}_{k'})}$$

- **M step:** maximization of the expectation of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{n_k^{[r]}}{n} \text{ and } \alpha_{kb}^{(h)[r+1]} = \begin{cases} \frac{n_{kb}^{(h)[r]}}{n_k^{[r]}} & \text{if } (1 \leq h \leq \ell_{kb}) \\ \frac{1 - \sum_{h'=1}^{\ell_{kj}} \alpha_{kb}^{(h')[r+1]}}{m^{\{b\}} - \ell_{kb}} & \text{otherwise} \end{cases}$$

Model selection (1/4)

$$(\hat{g}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\ell}}) = \underset{g,\boldsymbol{\sigma},\boldsymbol{\ell}}{\text{argmax}}\ p(g, \boldsymbol{\sigma}, \boldsymbol{\ell}|\mathbf{x})$$

1. $(\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\ell}}) = \underset{\boldsymbol{\sigma},\boldsymbol{\ell}}{\text{argmax}}\ p(\boldsymbol{\sigma}, \boldsymbol{\ell}|\mathbf{x}, g)$
2. $\hat{g} = \arg\max_g \text{BIC}(\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\ell}})$

Gibbs sampler

This algorithm has $p(\boldsymbol{\sigma}, \boldsymbol{\ell}|g, \mathbf{x})$ as marginal stationary distribution. Starting from an initial value $(\boldsymbol{\sigma}^{[0]}, \boldsymbol{\ell}^{[0]})$, the iteration $[s]$ is written as

$$\begin{aligned}
\boldsymbol{\theta}^{[s+1]} &\sim \boldsymbol{\theta}|(\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}), \mathbf{x}, \mathbf{z}^{[s]}, g \\
\mathbf{z}^{[s+1]} &\sim \mathbf{z}|(\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}), \mathbf{x}, \boldsymbol{\theta}^{[s+1]}, g \\
(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1]}) &\sim \boldsymbol{\sigma}, \boldsymbol{\ell}|(\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}), \mathbf{x}, \mathbf{z}^{[s+1]}, g \quad \longrightarrow \text{MCMC1 algorithm}
\end{aligned}$$

- $p(\boldsymbol{\sigma})$ (g and $\boldsymbol{\sigma}$ independent) and $p(\boldsymbol{\ell}|g, \boldsymbol{\sigma})$ follow uniform distributions
- $p(g) = \frac{1}{g_{\text{max}}}$ for $g = 1, \ldots, g_{\text{max}}$
- Poor informative priors on $\boldsymbol{\theta}$

Model selection (2/4)

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1]}) \sim \boldsymbol{\sigma}, \boldsymbol{\ell} | \boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1],g}$$

MCMC1 algorithm

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}) \sim \boldsymbol{\sigma}, \boldsymbol{\ell} | \boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1]}, g \quad \longrightarrow \text{MH algorithm}$$

$$\boldsymbol{\ell}^{[s+1]} \sim \boldsymbol{\ell} | \boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}, g \quad \longrightarrow \text{MCMC2 algorithm}$$

Model selection (3/4)

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}) \sim \boldsymbol{\sigma}, \boldsymbol{\ell} | \boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}, \mathbf{x}, \mathbf{z}^{[s+1]}, g$$

Metropolis-Hastings algorithm

- **Proposal distribution:**

$$(\boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star}) \sim q((\boldsymbol{\sigma}, \boldsymbol{\ell}); (\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}))$$

- **Acceptance distribution:**

$$\lambda^{[s]} = \min \left\{ \frac{p(\mathbf{x}, \mathbf{z}^{[s+1]} | (\boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star}))}{p(\mathbf{x}, \mathbf{z}^{[s+1]} | (\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}))} \frac{q((\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]})); (\boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star}))}{q((\boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star}); (\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}))}; 1 \right\}.$$

$$(\boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}) = \begin{cases} (\boldsymbol{\sigma}^{\star}, \boldsymbol{\ell}^{\star}) & \text{with a probability } \lambda^{[s]} \\ (\boldsymbol{\sigma}^{[s]}, \boldsymbol{\ell}^{[s]}) & \text{with a probability } 1 - \lambda^{[s]}. \end{cases}$$

See next slide for computation of $p(\mathbf{x}, \mathbf{z} | (\boldsymbol{\sigma}, \boldsymbol{\ell}))$

Model selection (4/4)

$$\boldsymbol{\ell}^{[s+1]} \sim \boldsymbol{\ell} | \boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}, g$$

MCMC2 algorithm

This step allows us to increase or decrease the mode number of each block by one at each iteration:

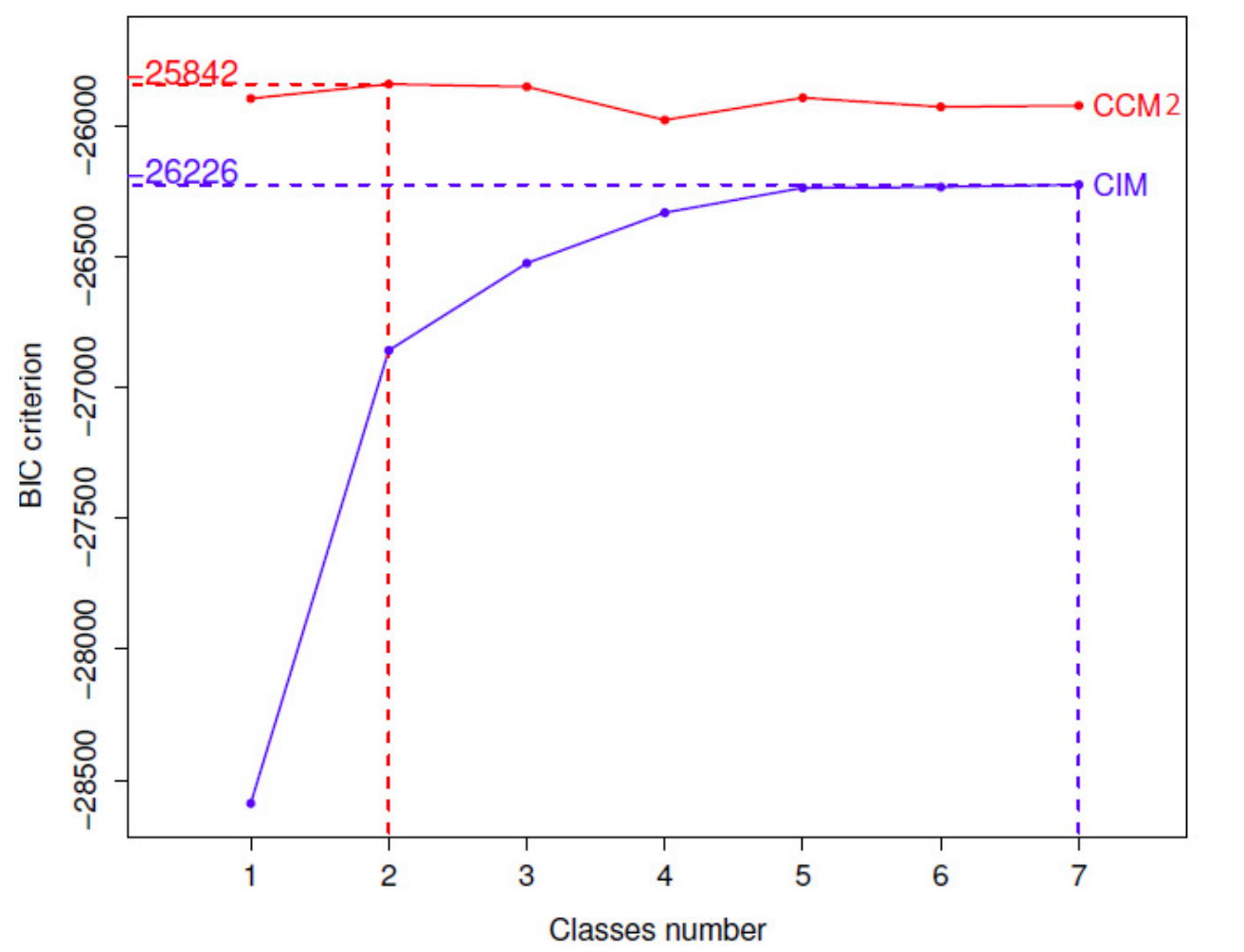
$$p(\ell_{kb} | \boldsymbol{\sigma}^{[s+1]}, \boldsymbol{\ell}^{[s+1/2]}, \mathbf{x}, \mathbf{z}^{[s+1]}) \propto \begin{cases} p(\mathbf{x}^{\{b\}} | \mathbf{z}^{[s+1]}, \ell_{kb}) & \text{if } |\ell_{kb} - \ell_{kb}^{[s+1/2]}| < 2 \\ & \text{and } \ell_{kb} \notin \{0, m^{\{b\}}\}. \\ 0 & \text{otherwise} \end{cases}$$

with

$$p(\mathbf{x}^{\{b\}} | \mathbf{z}, \ell_{kb}) \approx \left(\frac{1}{m^{\{b\}} - \ell_{kb}} \right)^{\bar{\mathrm{n}}_{kb}^{\ell_{kb}}} \prod_{h=1}^{\ell_{kb}} \frac{Bi\left(\frac{1}{m^{\{b\}} - h + 1}; \mathrm{n}_{kb}^{(h)} + 1; \bar{\mathrm{n}}_{kb}^{h} + 1 \right)}{m^{\{b\}} - h},$$

where $Bi(x; a, b) = B(1; a, b) - B(x; a, b)$ and where $B(x; a, b)$ is the incomplete beta function defined by $B(x; a, b) = \int_0^x w^a (1 - w)^b dw$.

Calves (1/2)



CCM2 $>$ CCM1 $>$ CIM

Calves (2/2)

Classes interpretation

- Class 1:
  - Represents 56% of calves.
  - The less protected ones (preventive treatment).
- Class 2:
  - Represents 44% of calves.
  - The most protected ones (preventive treatment).

Discriminative variables

- **Aptitude** is not discriminative (same modes and probabilities in both classes).
- **Treatment Omphalite** very discriminative:
  - Class 1: no treatment (0.92) .
  - Class 2: preventive treatment (0.93).

Dentistry (1/3)

g	1	2	3	4
CIM	-8766	-7511	**-7481**	-7503
CMM1	-7743	**-7473**	-7481	-7503
CMM2	-8294	-7492	**-7481**	-7503

CCM1 $>$ CCM2 $>$ CIM

Dentistry (2/3)

Number of variables: 5 Number of individuals: 3869
Number of modalities: 2 2 2 2 2
Class number: 2 log-likelihood: -7434.628 BIC: -7492.453
****

Mode number:

	de1-de2-de3	de4	de5
Class 2	5	1	1
Class 1	4	1	1

****

Tau index:

	de1-de2-de3	de4	de5
Class 2	0.8495717	0.5477190	0.8945463
Class 1	1.0000000	0.9798947	0.7185214

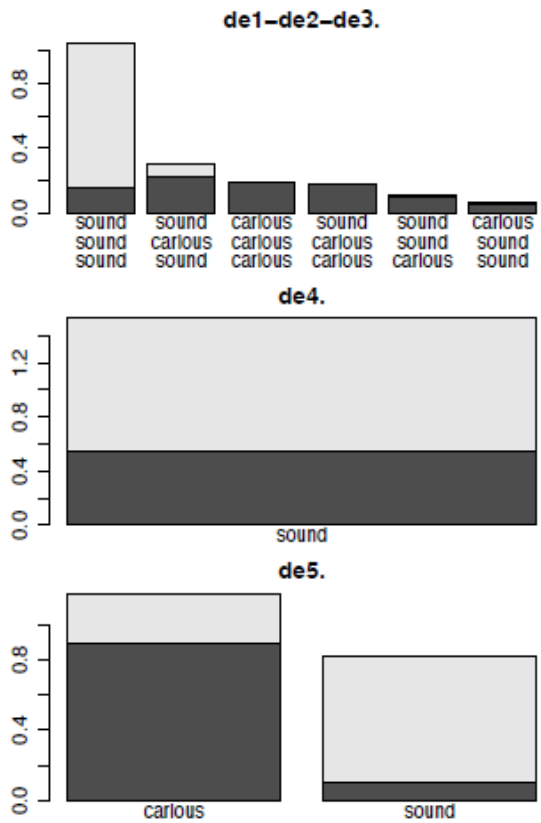
****

Kappa index:

	de1-de2-de3	de4	de5
Class 2	0.7142857	1	1
Class 1	0.5714286	1	1

Dentistry (3/3)

The majority class (displayed in gray) is mainly composed with the sound diagnoses. The second class (displayed in black) is composed with teeth diagnosed as carious by some dentists especially the fifth. Note that the dentist 4 mainly diagnoses the teeth as sound since its corresponding variable has a mode in this location for both classes.



Packages

- CIM: http://cran.r-project.org/web/packages/Rmixmod/index.html
- CCM1: http://r-forge.r-project.org/projects/clustericat/
- CCM2: https://r-forge.r-project.org/projects/comodes/