Modèle de mélange de copules Gaussiennes pour la classification des données hétérogènes.

Matthieu Marbac [1] & Christophe Biernacki[2] & Vincent Vandewalle [3].

[1] Inria Lille & DGA
[2] Université Lille 1 & CNRS & Inria Lille
[3] Université Lille 2 & Inria Lille.

Jeudi 29 août 2013.

Cadre de travail

Objectif : classification non supervisée de données hétérogènes.

Outil : **modèle de mélange**.

$$\overbrace{\begin{bmatrix} 2.4 & -9.2 & non & 3 & grand & \dots \\ 5.6 & -8.6 & oui & 5 & petit & \dots \\ 8.2 & 4.0 & non & 4 & petit & \dots \\ -2.0 & 2.6 & non & 6 & moyen & \dots \\ -1.6 & 9.6 & oui & 6 & grand & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}^{\boldsymbol{x}} \quad \overbrace{\begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ \vdots \end{bmatrix}}^{z}$$

- $\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{C}}, \boldsymbol{x}^{\mathrm{D}})$ vecteur de d variables ;
- $\boldsymbol{x}^{\mathrm{C}}$ vecteur de d_{C} variables continues ;
- $\boldsymbol{x}^{\mathrm{D}}$ vecteur de d_{D} variables discrètes (entières, binaires ou ordinales).

Cadre de travail

Objectif : classification non supervisée de données hétérogènes.

Outil : **modèle de mélange**.

$$\overbrace{\begin{bmatrix} 2.4 & -9.2 & non & 3 & grand & \dots \\ 5.6 & -8.6 & oui & 5 & petit & \dots \\ 8.2 & 4.0 & non & 4 & petit & \dots \\ -2.0 & 2.6 & non & 6 & moyen & \dots \\ -1.6 & 9.6 & oui & 6 & grand & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}^{\boldsymbol{x}} \quad \overbrace{\begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ \vdots \end{bmatrix}}^{z}$$

- $\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{C}}, \boldsymbol{x}^{\mathrm{D}})$ vecteur de d variables ;
- $\boldsymbol{x}^{\mathrm{C}}$ vecteur de d_{C} variables continues ;
- $\boldsymbol{x}^{\mathrm{D}}$ vecteur de d_{D} variables discrètes (entières, binaires ou ordinales).

Cadre de travail

Objectif : classification non supervisée de données hétérogènes.

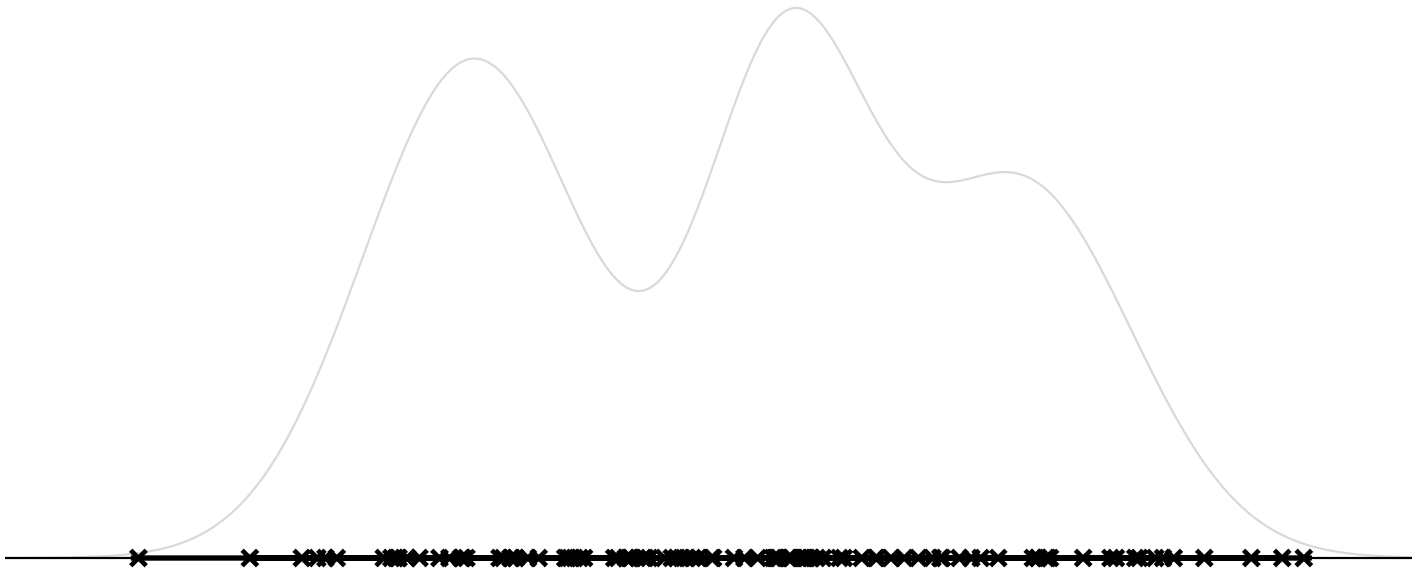
Outil : **modèle de mélange**.

$$\overbrace{\begin{bmatrix} 2.4 & -9.2 & non & 3 & grand & \dots \\ 5.6 & -8.6 & oui & 5 & petit & \dots \\ 8.2 & 4.0 & non & 4 & petit & \dots \\ -2.0 & 2.6 & non & 6 & moyen & \dots \\ -1.6 & 9.6 & oui & 6 & grand & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}}^{\boldsymbol{x}} \quad \overbrace{\begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ \vdots \end{bmatrix}}^{\boldsymbol{z}}$$

- $\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{C}}, \boldsymbol{x}^{\mathrm{D}})$ vecteur de d variables ;
- $\boldsymbol{x}^{\mathrm{C}}$ vecteur de d_{C} variables continues ;
- $\boldsymbol{x}^{\mathrm{D}}$ vecteur de d_{D} variables discrètes (entières, binaires ou ordinales).

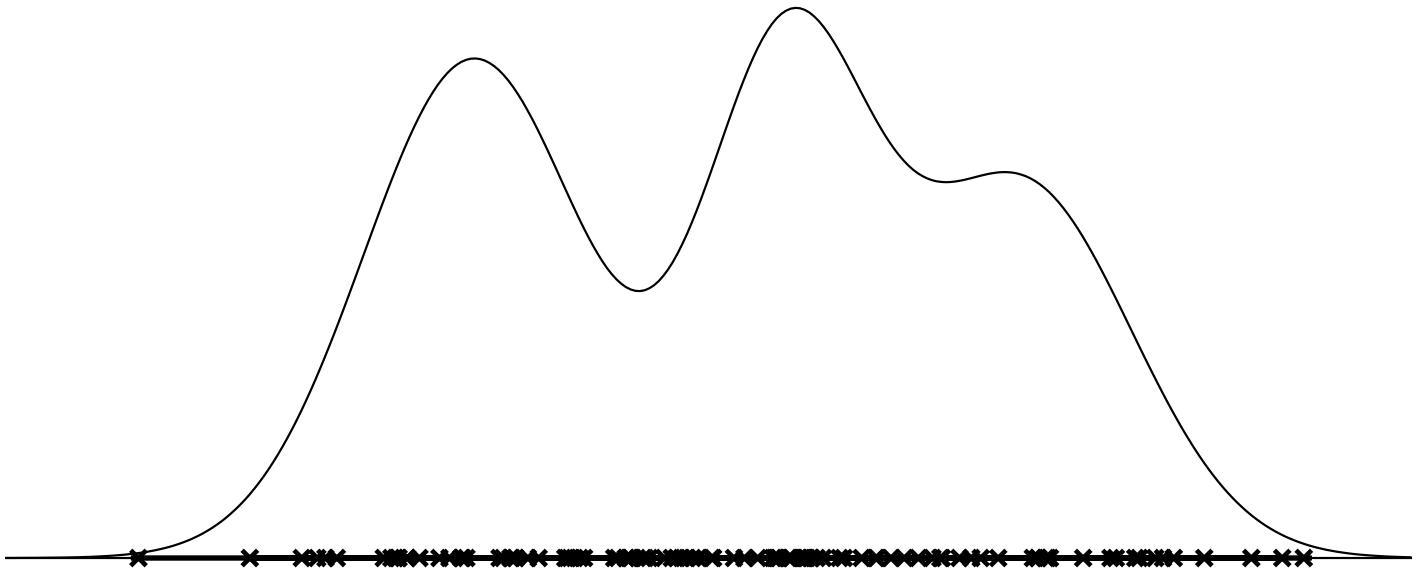
Modèles de mélange

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x};\boldsymbol{\theta}_k).$$



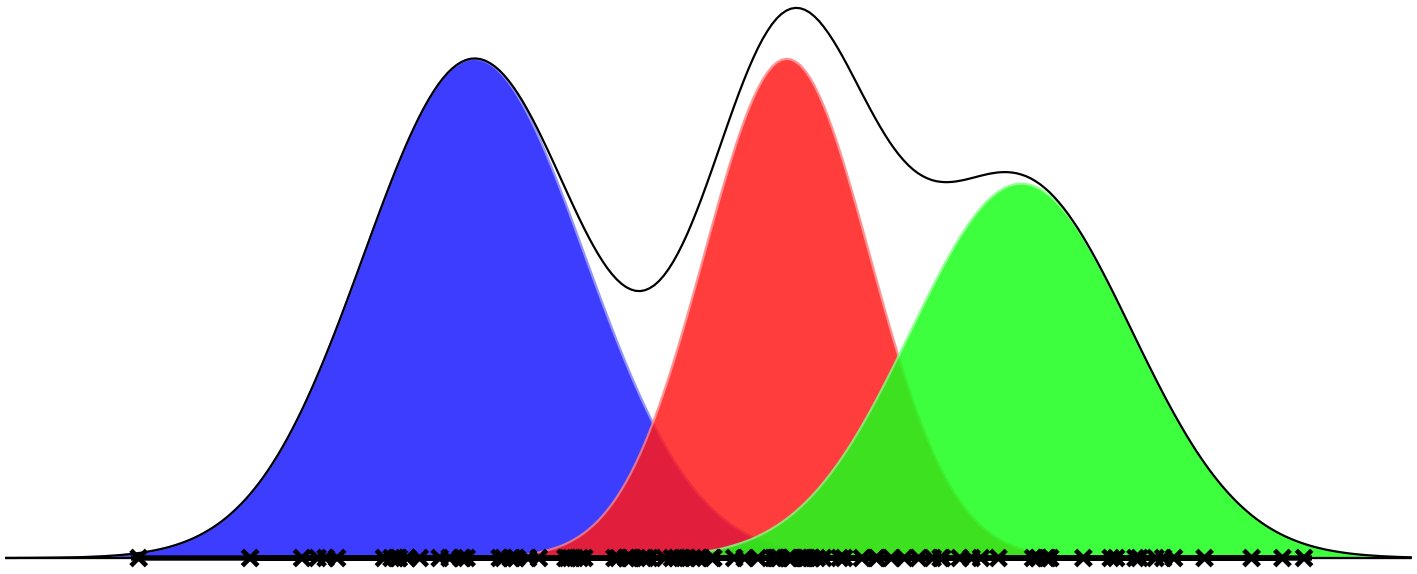
Modèles de mélange

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x};\boldsymbol{\theta}_k).$$



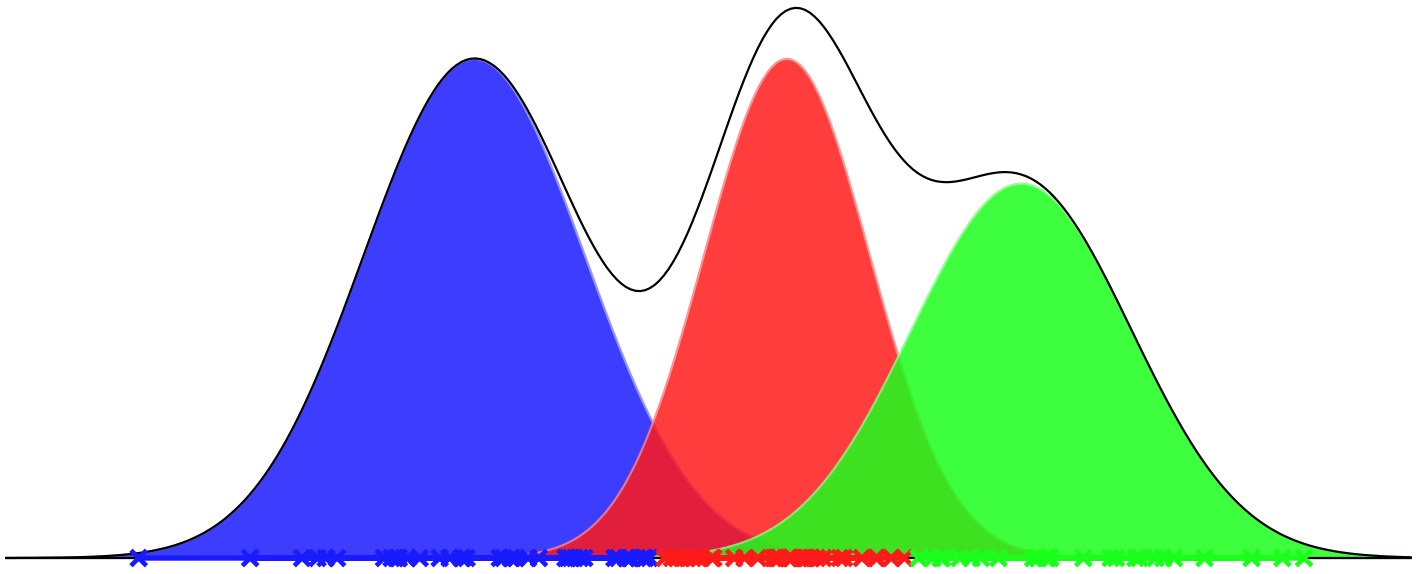
Modèles de mélange

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x};\boldsymbol{\theta}_k).$$



Modèles de mélange

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x};\boldsymbol{\theta}_k).$$



Modèles de mélange

Nature	Composante
Continue	Gaussienne multivariée
Entière	Poisson multivarié
Ordinale	Multinomiales
Binaire	Bernoulli
Hétérogène	**? ? ?**

Objectifs pour le cas hétérogène :

- composantes interprétables ;
- conservation des distributions homogènes ;
- prise en compte des dépendances intra-classe.

Copule (définition)

C'est une fdr définie sur $[0,1]^d$ avec des marges uniformes sur $[0,1]$.

Modèles de mélange

Nature	Composante
Continue	Gaussienne multivariée
Entière	Poisson multivarié
Ordinale	Multinomiales
Binaire	Bernoulli
Hétérogène	**? ? ?**

Objectifs pour le cas hétérogène :

- composantes interprétables ;
- conservation des distributions homogènes ;
- prise en compte des dépendances intra-classe.

Copule (définition)

C'est une fdr définie sur $[0,1]^d$ avec des marges uniformes sur $[0,1]$.

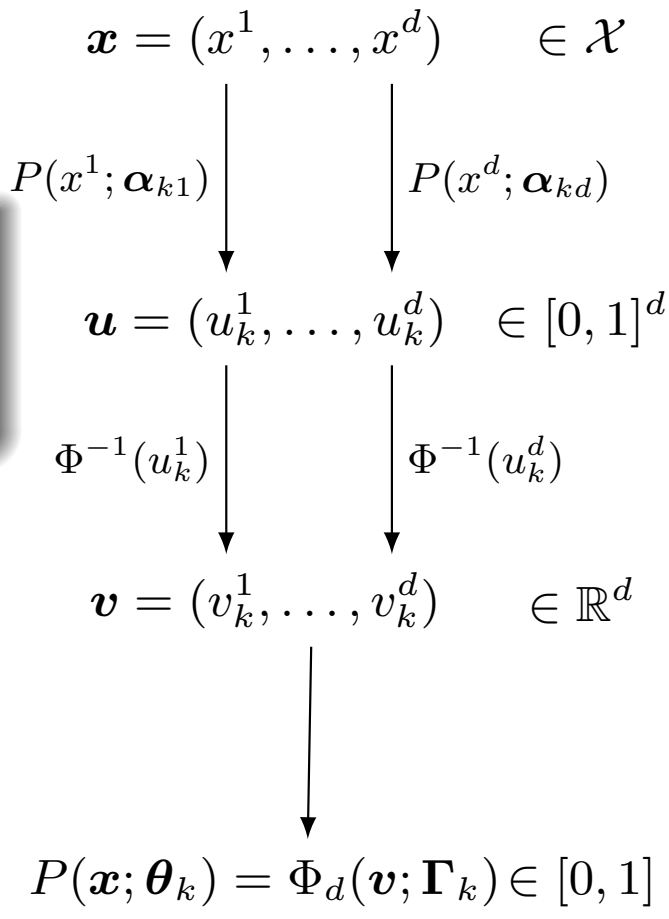
Mélange de copules gaussiennes

Copule Gaussienne (définition)

La fdr de la composante k est $P(\boldsymbol{x};\boldsymbol{\theta}_k)$ avec

$$P(\boldsymbol{x};\boldsymbol{\theta}_k) = \Phi_d(\Phi_1^{-1}(u_k^1),\ldots,\Phi_1^{-1}(u_k^d);\boldsymbol{\Gamma}_k).$$

- $\Phi_d(.;.)$ fdr de $\mathcal{N}(\mathbf{0},\boldsymbol{\Gamma}_k)$;
- $\boldsymbol{\Gamma}_k$ matrice des corrélations ;
- $P(x^j;\boldsymbol{\alpha}_{kj})$ fdr de x^j pour la composante k.



Mélange de copules gaussiennes

Modèle génératif

Échantillonnage de l'appartenance aux classes

$$\boldsymbol{z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$$

Échantillonnage des variables Gaussiennes

$$\boldsymbol{y} \sim \phi_d(\mathbf{0}; \boldsymbol{\Gamma}_k)$$

Calcul des variables observées

$$\boldsymbol{x} = \psi(\boldsymbol{y}; \boldsymbol{\alpha}_k)$$

où $\psi(\boldsymbol{y}; \boldsymbol{\alpha}_k) = (\psi(y^1; \boldsymbol{\alpha}_{k1}), \ldots, \psi(y^d; \boldsymbol{\alpha}_{kd}))$, $\psi(y^j; \boldsymbol{\alpha}_{kj}) = P^{-1}(\Phi_1(y^j); \boldsymbol{\alpha}_{kj})$.

- si x^j continue alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est bijective ;
- si x^j discret alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est surjective.

Mélange de copules gaussiennes

Modèle génératif

Échantillonnage de l'appartenance aux classes

$$\boldsymbol{z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$$

Échantillonnage des variables Gaussiennes

$$\boldsymbol{y} \sim \phi_d(\mathbf{0}; \boldsymbol{\Gamma}_k)$$

Calcul des variables observées

$$\boldsymbol{x} = \psi(\boldsymbol{y}; \boldsymbol{\alpha}_k)$$

où $\psi(\boldsymbol{y}; \boldsymbol{\alpha}_k) = (\psi(y^1; \boldsymbol{\alpha}_{k1}), \ldots, \psi(y^d; \boldsymbol{\alpha}_{kd}))$, $\psi(y^j; \boldsymbol{\alpha}_{kj}) = P^{-1}(\Phi_1(y^j); \boldsymbol{\alpha}_{kj})$.

- si x^j continue alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est bijective ;
- si x^j discret alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est surjective.

Mélange de copules gaussiennes

Modèle génératif

Échantillonnage de l'appartenance aux classes

$$\boldsymbol{z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$$

Échantillonnage des variables Gaussiennes

$$\boldsymbol{y} \sim \phi_d(\mathbf{0}; \boldsymbol{\Gamma}_k)$$

Calcul des variables observées

$$\boldsymbol{x} = \psi(\boldsymbol{y}; \boldsymbol{\alpha}_k)$$

où $\psi(\boldsymbol{y}; \boldsymbol{\alpha}_k) = (\psi(y^1; \boldsymbol{\alpha}_{k1}), \ldots, \psi(y^d; \boldsymbol{\alpha}_{kd}))$, $\psi(y^j; \boldsymbol{\alpha}_{kj}) = P^{-1}(\Phi_1(y^j); \boldsymbol{\alpha}_{kj})$.

- si x^j continue alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est bijective ;
- si x^j discret alors $x^j = \psi(y^j; \boldsymbol{\alpha}_{kj})$ est surjective.

Mélange de copules gaussiennes

Une copule gaussienne par composante : liaison par les fdr.

$$P(\boldsymbol{x};\boldsymbol{\theta}_k) = \Phi_d(\Phi_1^{-1}(u_k^1),\ldots,\Phi_1^{-1}(u_k^d);\boldsymbol{\Gamma}_k). \quad (1)$$

Mélange de copules gaussiennes : fonction de densité

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x};\boldsymbol{\theta}_k). \quad (2)$$

$$p(\boldsymbol{x};\boldsymbol{\theta}_k) = \int_{\mathcal{D}(\boldsymbol{x}^{\mathrm{D}};\boldsymbol{\alpha}_k)} \phi_d(\boldsymbol{y}_k;\boldsymbol{\Gamma}_k) d\boldsymbol{y}_k^{\mathrm{D}} \prod_{j=1}^{d_{\mathrm{C}}} \frac{p(x^j;\boldsymbol{\alpha}_{kj})}{\phi_1(y^j)}. \quad (3)$$

- $\mathcal{D}(\boldsymbol{x}^{\mathrm{D}};\boldsymbol{\alpha}_k) = \otimes_{j=1+d_{\mathrm{D}}}^{d} \mathcal{D}(x^j;\boldsymbol{\alpha}_{kj})$;
- $\mathcal{D}(x^j;\boldsymbol{\alpha}_{kj}) = \{y_k^j : \ x^j = \psi(y_k^j;\boldsymbol{\alpha}_{kj})\}$;
- $\boldsymbol{y}_k = (\boldsymbol{y}_k^{\mathrm{C}}, \boldsymbol{y}_k^{\mathrm{D}})$;
- $\boldsymbol{y}_k^{\mathrm{D}} = (y_k^j; j = d_{\mathrm{C}}+1,\ldots,d)$.

Mélange de copules gaussiennes

Remarque 1 : visualisation

Les variables latentes gaussiennes permettent une visualisation des individus par classe par une ACP.

Remarque 2 : maximum de vraisemblance

SEM avec la variable latente $\boldsymbol{z}$ complexe car :

- calcul de $\int_{\mathcal{D}(\boldsymbol{x}^{\mathrm{D}};\boldsymbol{\alpha}_k)} \phi_d(\boldsymbol{y}_k;\boldsymbol{\Gamma}_k)d\boldsymbol{y}_k^{\mathrm{D}}$ couteux ;
- maximisation de l'espérance de la vraisemblance complétée en $\boldsymbol{z}$ très difficile.

Remarque 3 : maximum de vraisemblance

SEM avec les variables latentes $(\boldsymbol{z}, \boldsymbol{y})$ impossible (états absorbants).

Estimation des paramètres

Algorithme de Gibbs pour échantillonner selon $p(\boldsymbol{\theta}|\mathbf{x})$

Échantillonneur de Gibbs

$$
\begin{aligned}
\mathbf{z}^{(r)} &\sim p(\mathbf{z}|\mathbf{x}, \mathbf{y}^{(r)}, \boldsymbol{\theta}^{(r)}) \\
\boldsymbol{\pi}^{(r+1)} &\sim p(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z}^{(r)}) \\
(\boldsymbol{\alpha}_{k\mathrm{C}}^{(r+1)}, \mathbf{y}_k^{\mathrm{C}(r+1)}) &\sim p(\boldsymbol{\alpha}_{k\mathrm{C}}, \mathbf{y}_k^{\mathrm{C}}|\mathbf{x}, \mathbf{z}^{(r)}) \\
(\boldsymbol{\alpha}_{k\mathrm{D}}^{(r+1)}, \mathbf{y}_k^{\mathrm{D}(r+1)}) &\sim p(\boldsymbol{\alpha}_{k\mathrm{D}}, \mathbf{y}_k^{\mathrm{D}}|\mathbf{x}, \mathbf{y}_k^{\mathrm{C}(r+1)}, \mathbf{y}_k^{\mathrm{D}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\Gamma}_k^{(r)}) \\
\boldsymbol{\Gamma}_k^{(r+1)} &\sim p(\boldsymbol{\Gamma}_k|\mathbf{y}^{(r+1)}, \mathbf{z}^{(r)}).
\end{aligned}
$$

Choix de modèle

L'objectif est de trouver le modèle maximisant :

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4}$$

Deux critères de choix de modèle

- Critère BIC qui approxime l'intégrale par :

$$\ln p(\mathbf{x}) \simeq \ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}) - \frac{\nu}{2}\ln n. \tag{5}$$

- Critère de Newton-Raftery qui utilise les sorties de l'échantillonneur de Gibbs :

$$p(\mathbf{x}) = \left[\frac{1}{\mathrm{R}}\sum_{r=1}^{\mathrm{R}}\frac{1}{p(\mathbf{x}|\boldsymbol{\theta}^{(r)})}\right]^{-1}. \tag{6}$$

Maladie du foie

Description du jeu de données

- 345 patients ;
- 5 variables continues relatives à des tests sanguins ;
- 1 variable entière décrivant le nombre de verres consommés par jour.

g	1	2	3	4	5	6
MMCG	-8040	**-7079**	-7093	-7137	-7113	-7261
MMDG	-8024	-7765	-7717	**-7712**	-7713	-7718

TABLE: Valeurs du critère BIC pour le modèle de mélange de copules Gaussiennes (MMCG) et pour le modèle de mélange de distributions Gaussiennes (MMDG).

Maladie du foie

classe	π_k	MOV	ALK	ALA	ASP	GGT	DRI
1	0.79	90.1 (4.3)	70.7 (16.5)	28.1 (16.0)	23.7 (8.4)	35.4 (28.0)	2.25
2	0.21	91.5 (4.1)	73.0 (20.4)	40.9 (25.1)	31.3 (10.9)	67.4 (60.0)	7.61

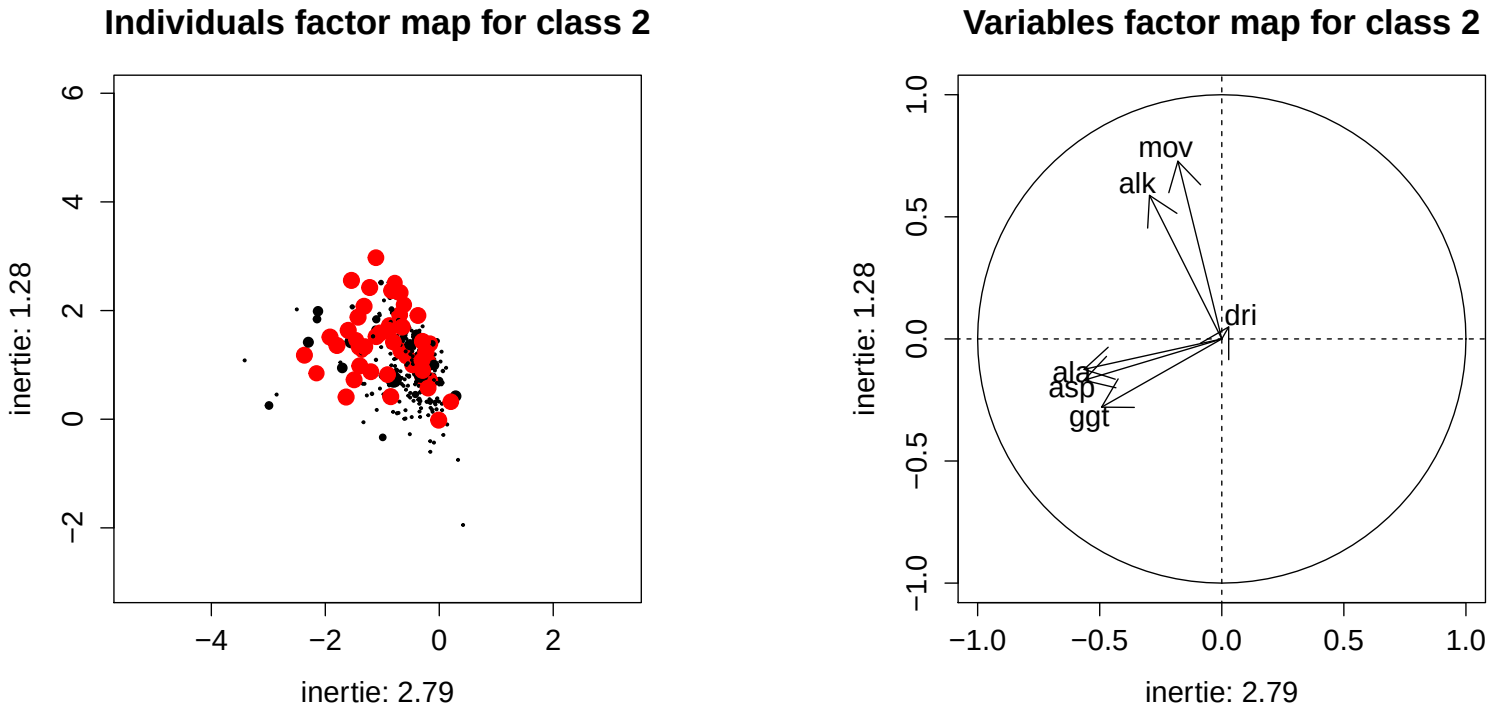
TABLE: Estimateurs du maximum de vraisemblances pour les marginales. Pour les variables continues, on donne la moyenne en plein et l'écart-type entre parenthèses.

Interprétation des classes

- classe 1
  - majoritaire,
  - peu ou pas de consommation d'alcool,
  - faibles valeurs et variances des tests sanguins ;
- classe 2
  - minoritaire,
  - forte consommation d'alcool,
  - mesures sanguines élevées avec grandes variances.

Maladie du foie

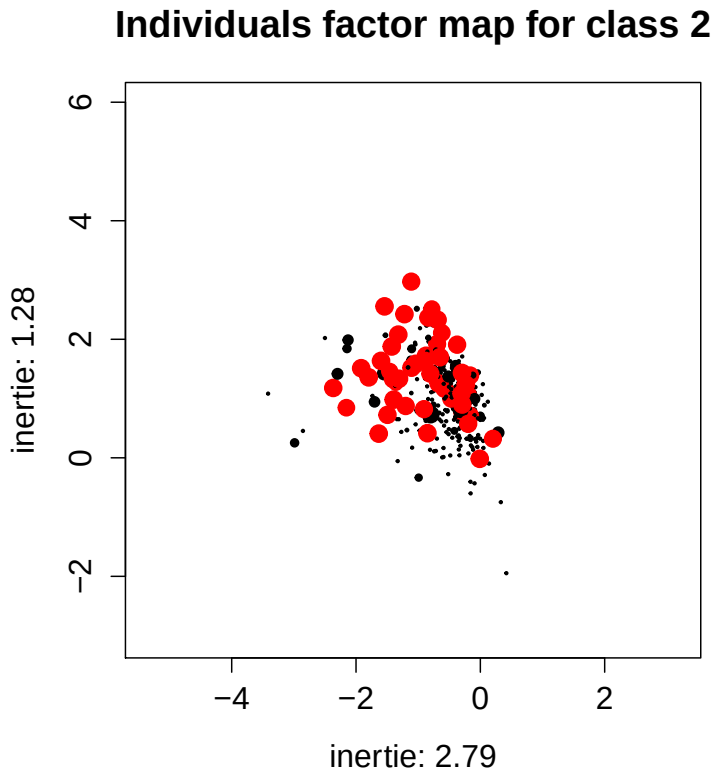
À partir des matrices de corrélations de chaque classe, on peut effectuer une ACP pour visualiser les données et les variables :



(a) (b)

FIGURE: Sorties de l'ACP effectuée sur la copule de la classe 2 : (a) individus dans le plan factoriel ; (b) variables dans le plan factoriel.

Conclusion



Échantillonnage de l'appartenance aux classes

$$\boldsymbol{z} \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$$

Échantillonnage des variables Gaussiennes

$$\boldsymbol{y} \sim \phi_d(\mathbf{0}; \boldsymbol{\Gamma}_k)$$

Calcul des variables observées

$$\boldsymbol{x} = \psi(\boldsymbol{y}; \boldsymbol{\alpha}_k)$$

Futures extensions

- prise en compte de données qualitatives non binaires ;
- critère de choix de modèle utilisant les sorties de l'échantillonneur de Gibbs plus stable que NR.