Classification de données mixtes par un modèle de mélange de copules Gaussiennes.

Matthieu Marbac

Travaux encadrés par
C. Biernacki et V. Vandewalle.

Université de Lille 1 & Inria Lille.

Vendredi 16 mai 2014.
Ateliers du GDR Mascot-Num.

Cadre de travail

Objectif : classification non supervisée de données mixtes.

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

- $\boldsymbol{x}_i$ vecteur de d variables.
- x_i^j est la j-ème variable.

Tous les modèles sont faux, certains sont utiles !

Cadre de travail

Objectif : classification non supervisée de données mixtes.

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

- $\boldsymbol{x}_i$ vecteur de d variables.
- x_i^j est la j-ème variable.

Tous les modèles sont faux, certains sont utiles !

Introduction aux modèles de mélanges.

Modèle de mélange

Hypothèse : les individus sont issus de g classes.

une classe = une distribution.

$\boldsymbol{Z}_i$ variable aléatoire indiquant la classe de l'individu i

$$\boldsymbol{Z}_i \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g),$$

π_k est la proportion de la classe k avec $0 < \pi_k \leq 1$ et $\sum_{k=1}^{g} \pi_k = 1$.

Challenge : estimer la réalisation $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ conditionnellement à la donnée observée $\boldsymbol{x}_i$.

$$z_{ik} = 1 \Leftrightarrow \boldsymbol{x}_i \text{ appartient à la classe } k.$$

La variable aléatoire $X_i|\boldsymbol{Z}_i = \boldsymbol{z}_i$ suit une distribution définie par la f.d.p. $f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)$. La f.d.p. du couple $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ est

$$f(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) = \prod_{k=1}^{g} \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)\right)^{z_{ik}},$$

où $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\alpha}_k; k = 1, \ldots, g)$. La f.d.p. des données observées $\boldsymbol{x}_i$ est

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k).$$

Modèle de mélange

Des composantes adaptées à la nature des variables

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k).$$

Nature des variables	$f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)$	Logiciels
Continues	gaussien	rmixmod mclust mixdist mplus
Entières	Poisson	mixdist
	binomiale	mixdist
Qualitatif	multinomiale	rmixmod
Mixtes	? ? ?	rmixmod MULTIMIX

package R, autre.

Modèle de mélange

Si $\boldsymbol{x}_i$ est un ensemble de variables *continues*, on utilise classiquement le modèle de *mélange gaussien* dont la f.d.p. est

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \text{ où } f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k).$$

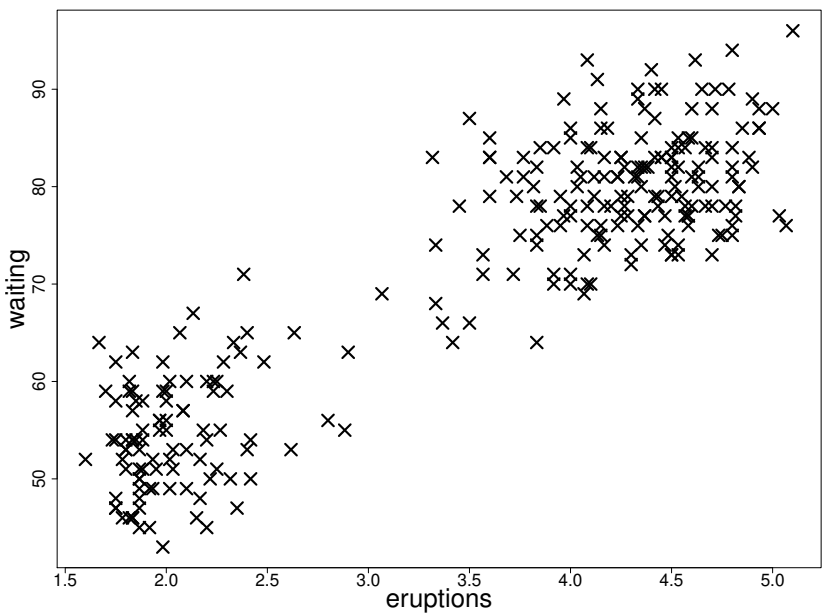


FIGURE: Jeu de données `faithful` (package MASS sous R).

Modèle de mélange

Si $\boldsymbol{x}_i$ est un ensemble de variables *continues*, on utilise classiquement le modèle de *mélange gaussien* dont la f.d.p. est

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \text{ où } f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k).$$

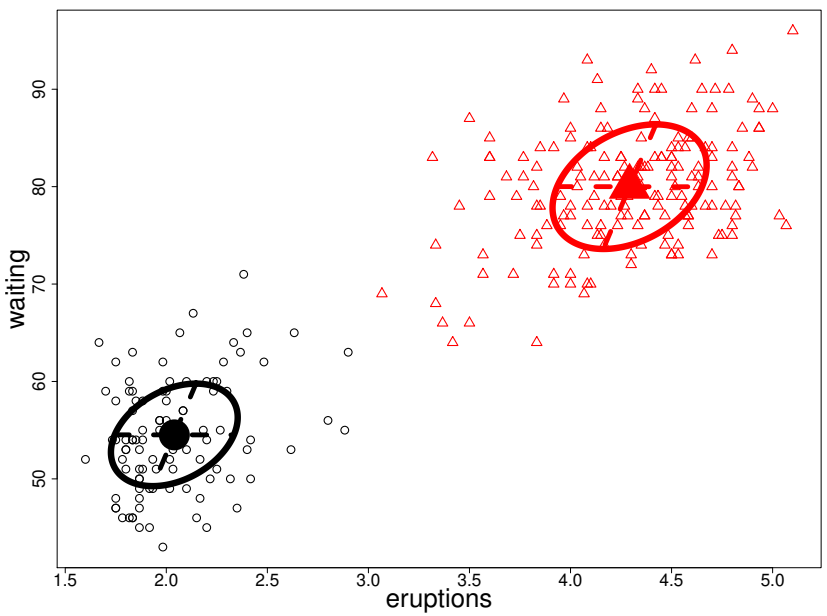


FIGURE: Partition obtenue avec un mélange gaussien à deux composantes.

Modèle de mélange

Estimation par maximum de vraisemblance (cadre fréquentiste)

À partir de l'échantillon i.i.d. $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$, on cherche argmax $L(\boldsymbol{\theta}; \mathbf{x})$

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \Big(\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \Big).$$

Algorithme EM

À l'itération (r), il s'écrit :

- **E step**

$$t_{ik}(\boldsymbol{\theta}^{(r)}) = \frac{\pi_k^{(r)} f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{(r)})}{\sum_{\ell=1}^{g} \pi_\ell^{(r)} f_\ell(\boldsymbol{x}_i; \boldsymbol{\alpha}_\ell^{(r)})}.$$

- **M step**

$$\boldsymbol{\theta}^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{k=1}^{n} \sum_{i=1}^{n} t_{ik}(\boldsymbol{\theta}^{(r)}) \ln \big(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \big).$$

Affectation des individus aux classes

$$\hat{z}_{ik} = 1 \text{ si } t_{ik}(\hat{\boldsymbol{\theta}}) = \underset{\ell=1,\ldots,g}{\operatorname{argmax}}\, t_{i\ell}(\hat{\boldsymbol{\theta}}).$$

Modèle de mélange

Estimation par maximum a posteriori (cadre Bayésien)

À partir de l'échantillon i.i.d. $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$, on cherche argmax $p(\boldsymbol{\theta}|\mathbf{x})$.

Échantillonneur de Gibbs
Sa distribution stationnaire est $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ et son itération (r) s'écrit :

$$\forall i = 1, \ldots, n \quad \boldsymbol{z}_i^{(r)} | \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)} \tag{1}$$

$$\boldsymbol{\pi}^{(r+1)} | \mathbf{x}, \mathbf{z}^{(r)} \tag{2}$$

$$\forall k = 1, \ldots, g \quad \boldsymbol{\alpha}_k^{(r+1)} | \mathbf{x}, \mathbf{z}^{(r)}. \tag{3}$$

Affectation des individus aux classes après $r_{\max}$ itérations de l'échantillonneur de Gibbs

$$\hat{z}_{ik} = 1 \text{ si } \sum_{r=1}^{r_{\max}} z_{ik}^{(r)} = \underset{\ell=1,\ldots,g}{\operatorname{argmax}} \sum_{r=1}^{r_{\max}} z_{i\ell}^{(r)}.$$

Modèle de mélange

Choix de modèle

Questions :

- Choix du nombre de classes : g ?
- Choix des composantes : $f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)$?

Problème :
La vraisemblance grandit naturellement avec le nombre de classes.

Outils :
Critères d'informations classiques : BIC, ICL...

$$BIC = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \frac{\nu}{2} \ln n,$$

ν est le nombre de paramètres du modèle.

Modèle de mélange

Bilan

- Une classe = une distribution de probabilité.
- Classification interprétable par les paramètres de composantes du mélanges (distributions classiques).
- Outils probabilistes disponibles pour les problèmes difficiles (choix de g).

Challenge des données mixtes

Approches naïves pour la classification de données mixtes

- **Conversion des variables ordinales/qualitatives en données continues**
  + méthode pour données qualitatives
  $\Longrightarrow$ fortes hypothèses peu réalistes (écarts identiques entre modalités, ordre).
- **Discrétisation des données continues**
  + méthode pour données qualitatives
  $\Longrightarrow$ pertes d'information.
- **ACM sur les données binaires et ordinales**
  + méthode pour données continues
  $\Longrightarrow$ interprétation et données "pseudo-continues".

Respect de la nature de chaque variables
$\Longrightarrow$ utilisation des modèles de mélange.

Modèle de mélange pour données mixtes

Problème des variables mixtes

Si $\boldsymbol{x}_i$ est un ensemble de variables mixtes il n'y a pas de distribution multivariée classique !

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k).$$

Challenge :
Déterminer les distributions multivariées $f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)$.

Cahier des charges :

- Distributions marginales des composantes sont classiques (gaussienne, Poisson, multinomiale...).
- Interprétation des dépendances pour $f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)$.

Deux solutions :

- Indépendance conditionnelle (implémentée dans rmixmod).
- Mélange de copules gaussiennes.

Modèle d'indépendance conditionnelle

Solution simple : modèle d'indépendance conditionnelle

- Les variables sont mutuellement indépendantes conditionnellement à la classe

$$\mathbb{P}(\boldsymbol{X}_i|\boldsymbol{Z}_i = \boldsymbol{z}_i) = \prod_{j=1}^{d} \mathbb{P}(X_i^j|\boldsymbol{Z}_i = \boldsymbol{z}_i).$$

- Le modèle à g composantes a pour f.d.p.

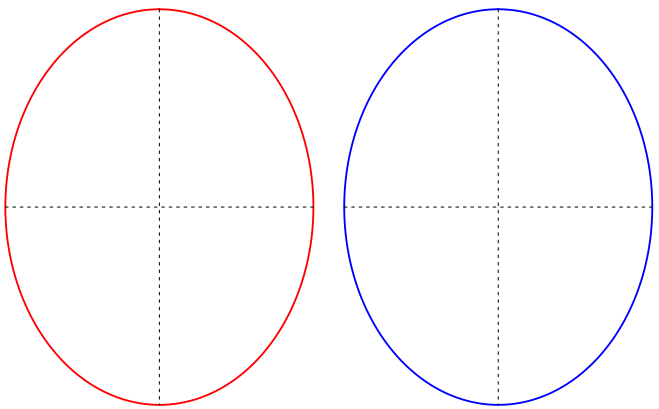
$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

- Facilement interprétable si les marginales de chaque classe ($f_{kj}(x^j;\boldsymbol{\alpha}_{kj})$) sont des distributions classiques (Gaussienne, Poisson, Multinomiale...).
- Efficace si peu de données.
- Implémenté dans `rmixmod`.

Solution simple : modèle d'indépendance conditionnelle

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

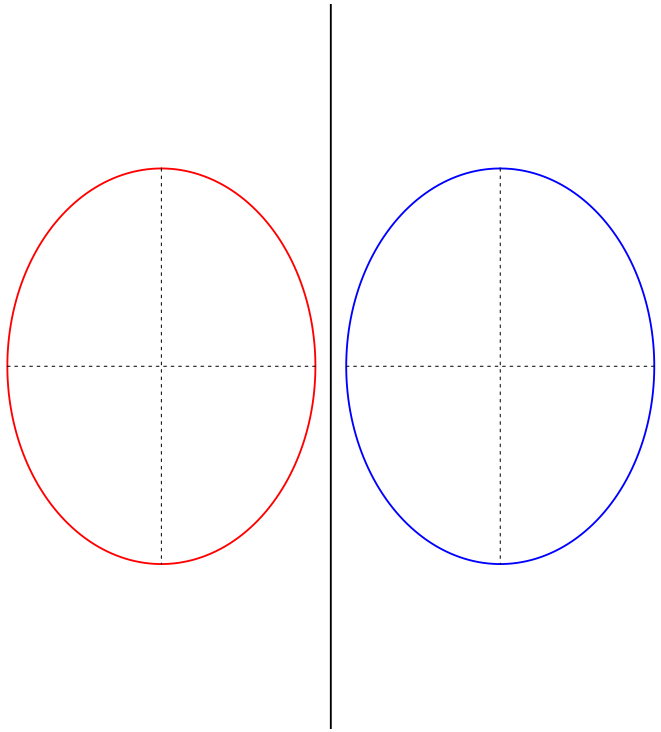
Ce modèle implique des biais lorsque les données sont corrélées dans une classe.



Solution simple : modèle d'indépendance conditionnelle

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

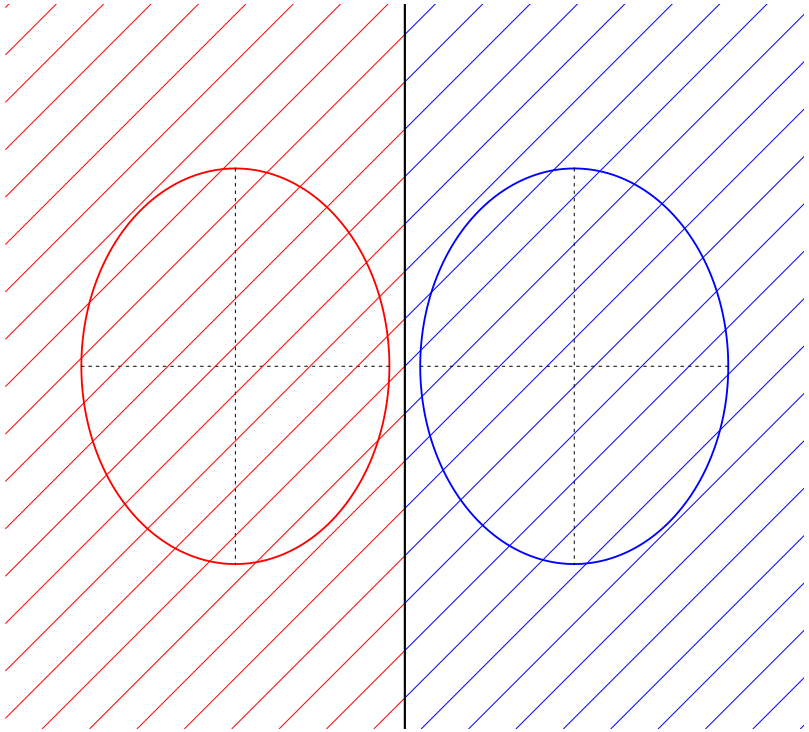
Ce modèle implique des biais lorsque les données sont corrélées dans une classe.



Solution simple : modèle d'indépendance conditionnelle

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

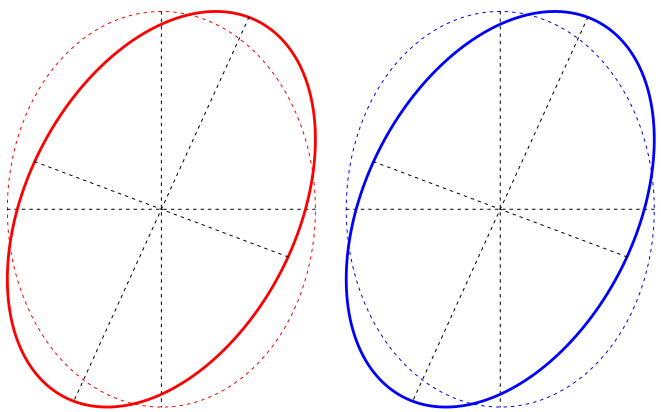
Ce modèle implique des biais lorsque les données sont corrélées dans une classe.



Solution simple : modèle d'indépendance conditionnelle

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

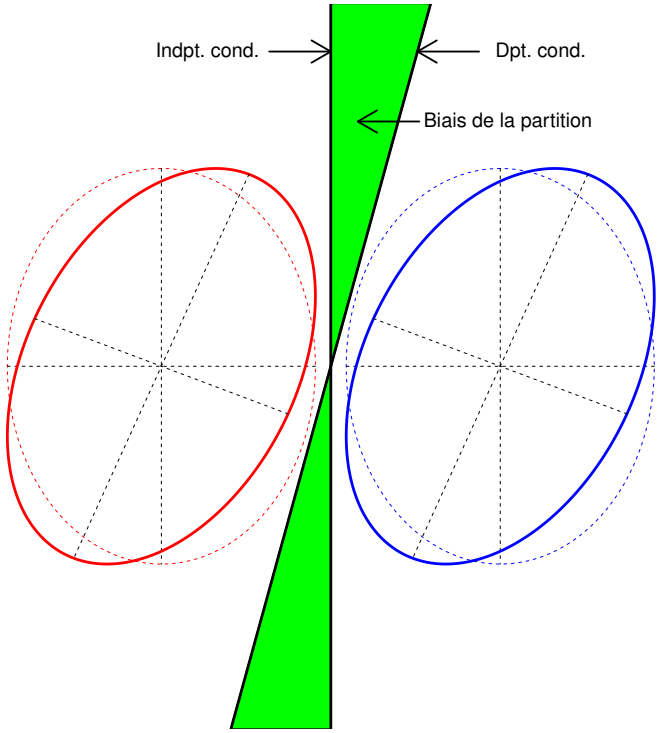
Ce modèle implique des biais lorsque les données sont corrélées dans une classe.



Solution simple : modèle d'indépendance conditionnelle

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) \quad \text{où} \quad f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_i^j;\boldsymbol{\alpha}_{kj}).$$

Ce modèle implique des biais lorsque les données sont corrélées dans une classe.



Modèle de mélange de copules gaussiennes

Objectif

Alternative au modèle d'indépendance conditionnelle

Objectifs pour le cas mixte :

- Composantes interprétables : $X_i^j | \boldsymbol{Z}_i = \boldsymbol{z}_i$ est une distribution classique.
- Prise en compte des dépendances intra-classe.

$\Longrightarrow$ Copules !

Propriété des copules

Définition d'une distribution multivariée en choisissant indépendamment :

- Les distributions marginales.
- La forme de la dépendance.

Copule

- Définition : une copule $C(u^1, \ldots, u^d)$ est une f.d.r. sur $[0,1]^d$ dont les marges sont uniformes sur $[0,1]$.
- Exemples :
  - Copules Gaussiennes

$$C(u^1, \ldots, u^d) = \Phi_d(\Phi_1^{-1}(u^1), \ldots, \Phi_1^{-1}(u^d); \mathbf{\Gamma}),$$

    où $\mathbf{\Gamma}$ est une matrice de corrélation et Φ_d f.d.r. de $\mathcal{N}_d(\mathbf{0}, \mathbf{\Gamma})$ et Φ_1 f.d.r. de $\mathcal{N}_1(0,1)$.
  - Copules de Student.
  - Copules archimédiennes (1 seul paramètre pour la dépendance).
- Théorème de Sklar, 1959
  Dans le cas continu, il existe une unique copule $C(u^1, \ldots, u^d)$ t.q.

$$F(x_i^1, \ldots, x_i^d) = C(F_1(x_i^1), \ldots, F_d(x_i^d)),$$

  $F_1, \ldots, F_d$ sont les f.d.r. marginales de $F(x_i^1, \ldots, x_i^d)$.

Pour résumer

- Copule $\Longrightarrow$ forme de la dépendance.
- Copule + Marginales $\Longrightarrow$ modèle multivarié.

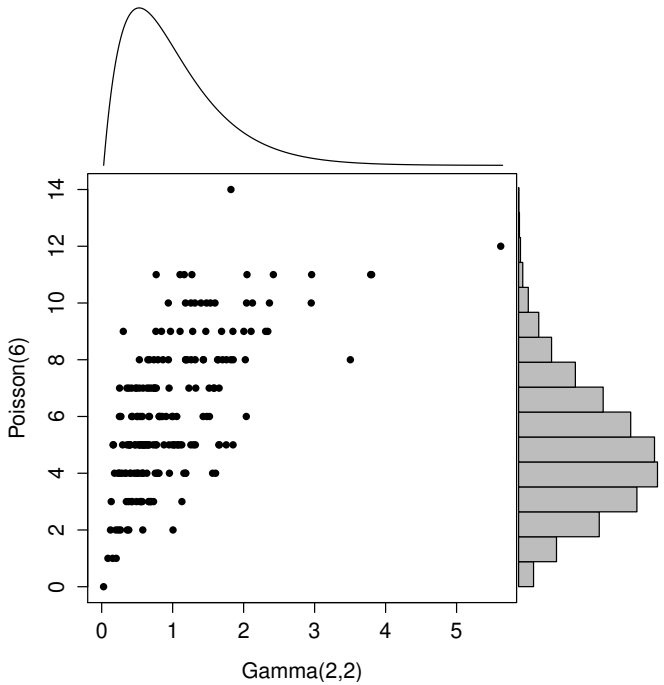
Copule Gaussienne : modèle génératif

- Copule Gaussienne :

$$\Phi_d(\Phi_1^{-1}(u_i^1), \ldots, \Phi_1^{-1}(u_i^d); \mathbf{\Gamma}),$$

où $u_i^j = F_j(x_i^j)$.

- Marginale j déterminées par $F_j(.)$.
- Exemple :

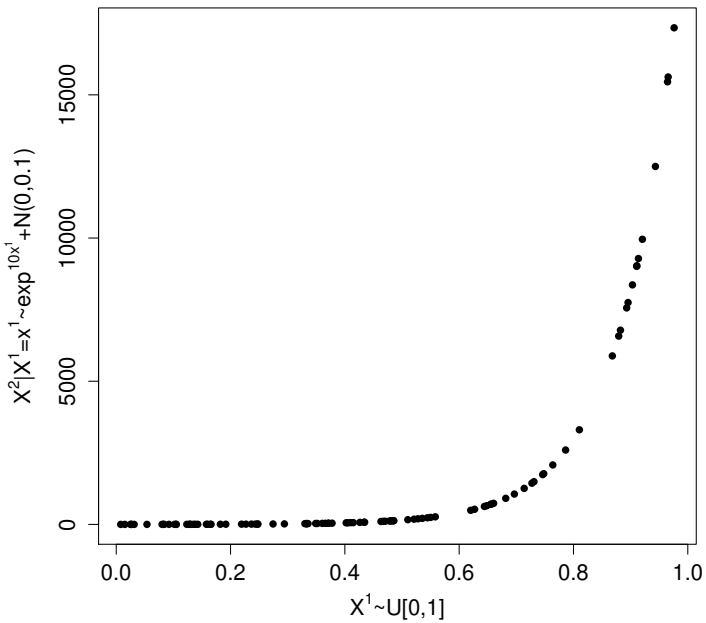


$$\begin{array}{l} \boldsymbol{y}_i = (y_i^1, \ldots, y_i^d) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{\Gamma}) \\ \quad \downarrow \Phi_1(y_i^1) \qquad \downarrow \Phi_1(y_i^d) \\ \boldsymbol{u}_i = (u_i^1, \ldots, u_i^d) \in [0,1]^d \\ \quad \downarrow F_1^{-1}(u_i^1) \qquad \downarrow F_d^{-1}(u_i^d) \\ \boldsymbol{x}_i = (x_i^1, \ldots, x_i^d) \end{array}$$

FIGURE: Copule gaussienne avec $\mathbf{\Gamma} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$.

Copule Gaussienne : propriétés

- Un coefficient de corrélation par couple de variables.
- Le coefficient de corrélation de la copule gaussienne est la borne supérieure des coefficients de corrélations obtenus par transformations monotones des variables (si continues).



$$\rho(X^1, X^2) = 0.69$$

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$$

FIGURE: Dépendance non linéaire entre 2 variables.

- Visualisation des données par composantes indépendantes (ACP sur données Gaussiennes).

Mélange de copules gaussiennes

- Une copule gaussienne par composante :

$$F_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \Phi_d(\Phi_1^{-1}(u_{ik}^1),\ldots,\Phi_1^{-1}(u_{ik}^d);\boldsymbol{\Gamma}_k),$$

où $u_{ik}^j = F_{kj}(x_i^j;\boldsymbol{\alpha}_{kj})$.

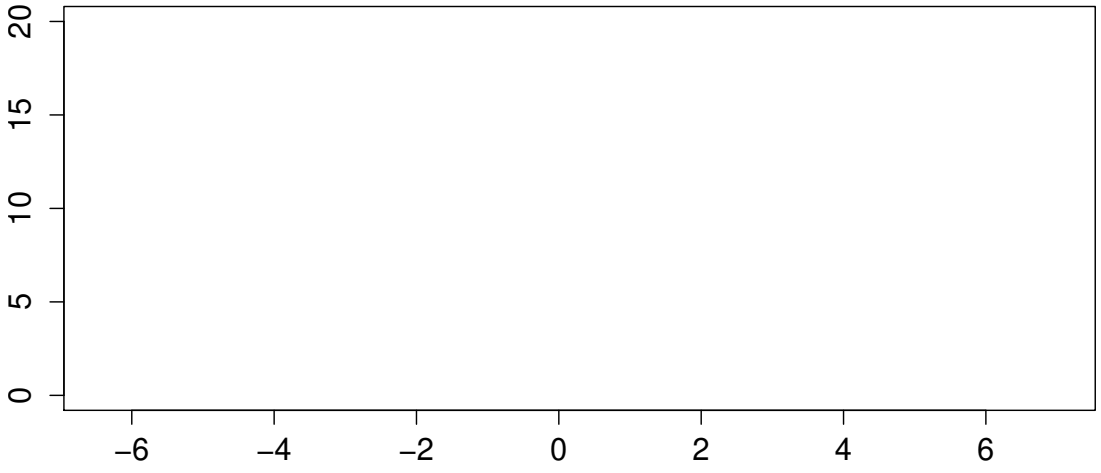
- Marges classiques pour les composantes du mélanges :
  - x_i^j continue : $F_{kj}(x_i^j;\boldsymbol{\alpha}_{kj})$ f.d.r. distribution gaussienne.
  - x_i^j entier : $F_{kj}(x_i^j;\boldsymbol{\alpha}_{kj})$ f.d.r. distribution de Poisson.
  - x_i^j ordinal : $F_{kj}(x_i^j;\boldsymbol{\alpha}_{kj})$ f.d.r. distribution multinomiale.
- F.d.p. de la composante k :

$$f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = f_k(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_k) f_k(\boldsymbol{x}_i^{\mathrm{D}}|\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_k) \tag{4}$$

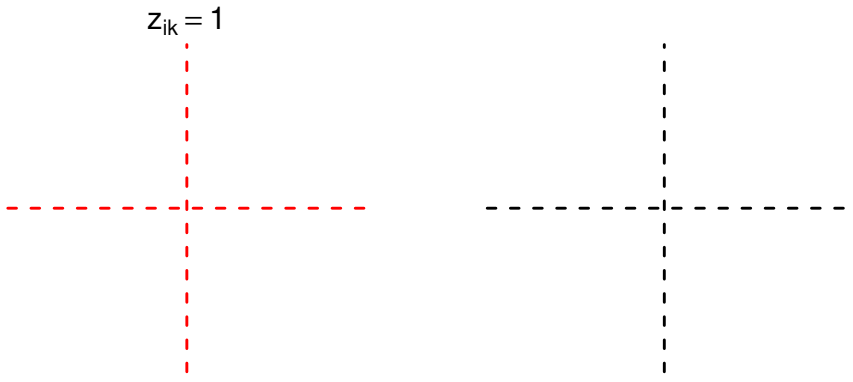
$$= f_k(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_k) \int_{\mathcal{S}_k(\boldsymbol{x}_i^{\mathrm{D}})} \phi(\boldsymbol{y}_i^{\mathrm{D}};\boldsymbol{\mu}_k^{\mathrm{D}},\boldsymbol{\Sigma}_k^{\mathrm{D}}) d\boldsymbol{y}_i^{\mathrm{D}}, \tag{5}$$

  - $f_k(\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_k)$ explicite (vbles continues).
  - $f_k(\boldsymbol{x}_i^{\mathrm{D}}|\boldsymbol{x}_i^{\mathrm{C}};\boldsymbol{\alpha}_k)$ peut être non explicite et difficile (vbles discrètes sachant les continues).

Modèle génératif du mélange de copules gaussiennes

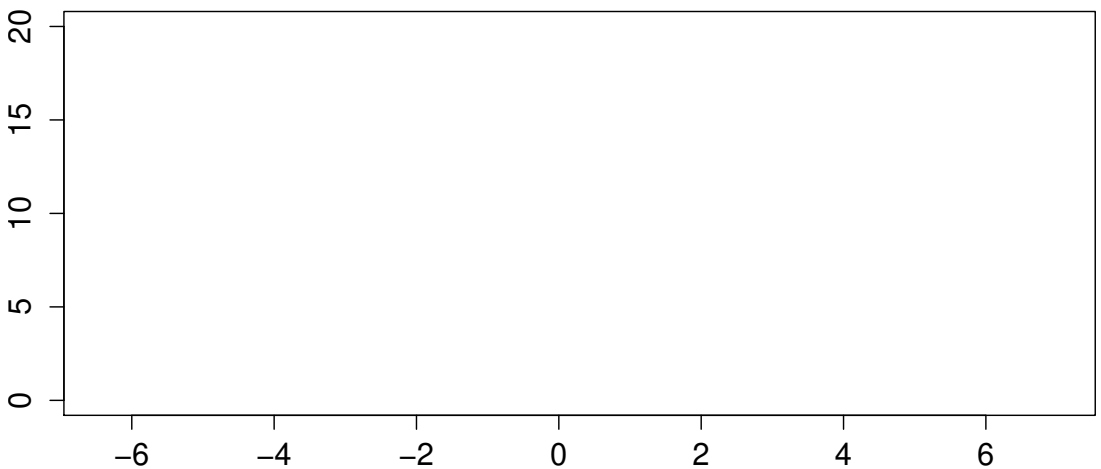


Modèle génératif du mélange de copules gaussiennes

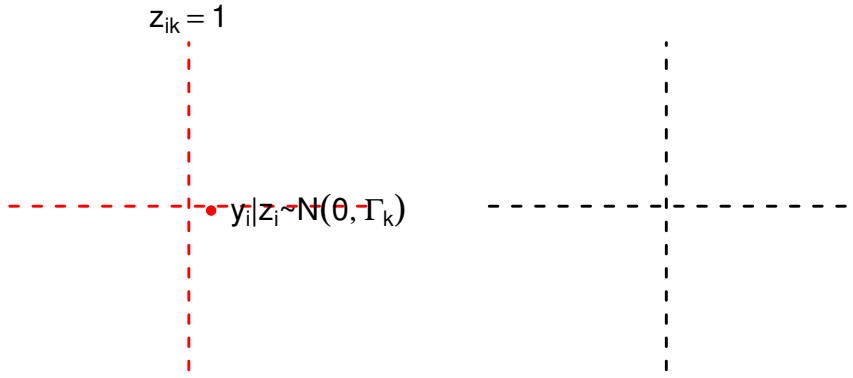


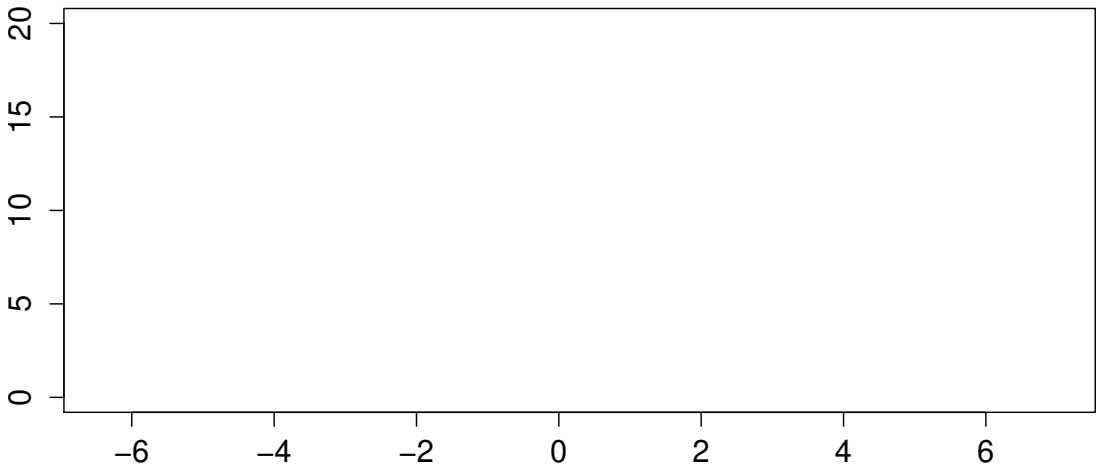
Échantillonnage de la classe

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



Modèle génératif du mélange de copules gaussiennes





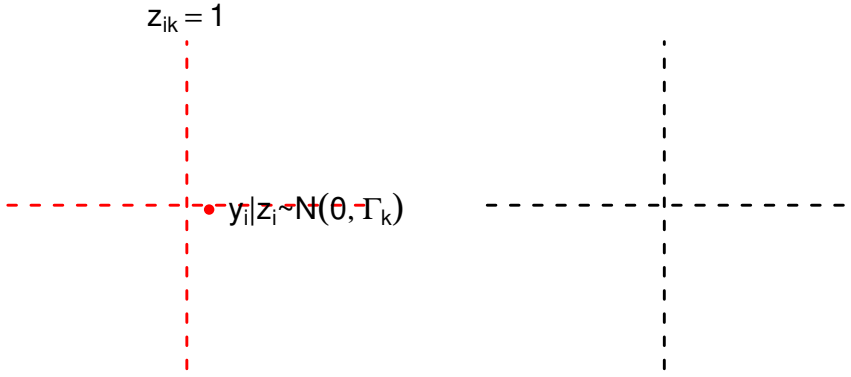
Échantillonnage de la classe

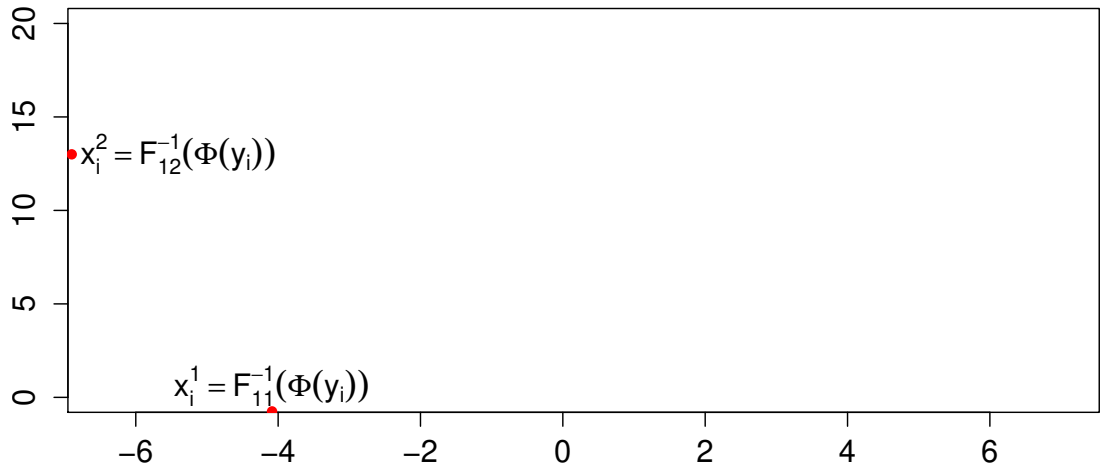
$$\boldsymbol{z}_i \sim \mathcal{M}_g(\pi_1, \ldots, \pi_g)$$

Échantillonnage Gaussien $\boldsymbol{y}_i|\boldsymbol{z}_{ik} = 1$

$$\boldsymbol{y}_i|\boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

Modèle génératif du mélange de copules gaussiennes





Échantillonnage de la classe

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

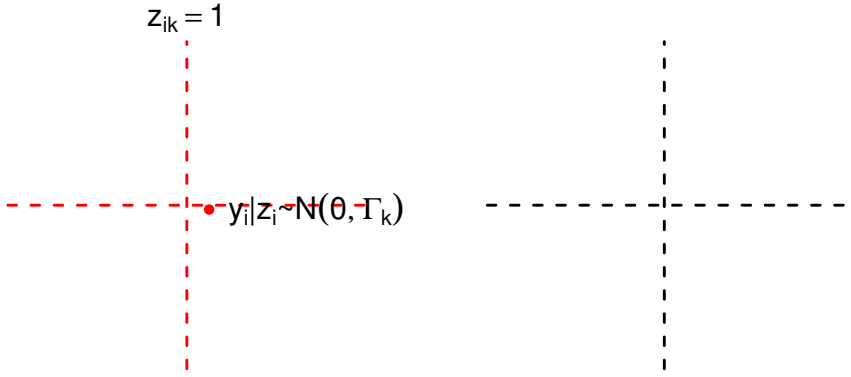
Échantillonnage Gaussien $\boldsymbol{y}_i|\boldsymbol{z}_{ik} = 1$

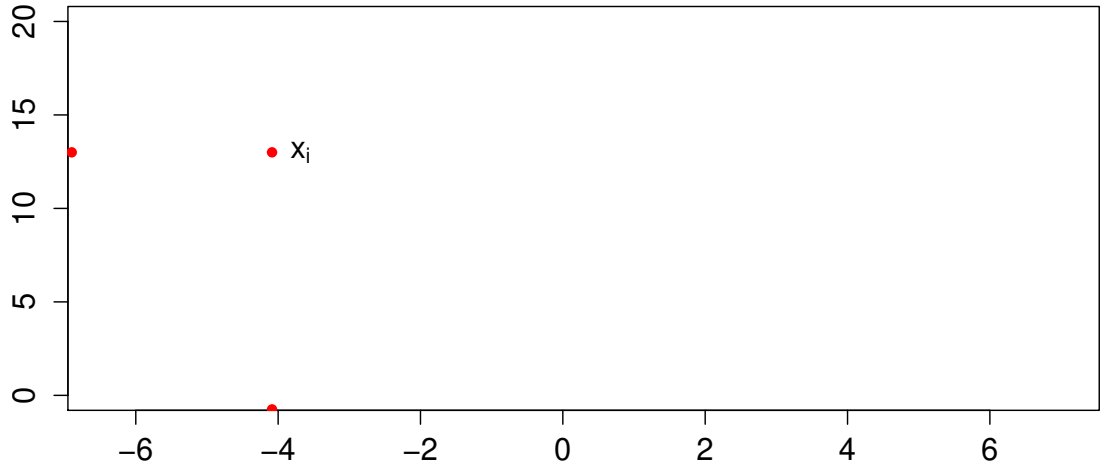
$$\boldsymbol{y}_i|\boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

Calcul de l'observation $\boldsymbol{x}_i|(\boldsymbol{z}_i, \boldsymbol{y}_i)$

$$x_i^j = F_{kj}^{-1}(\Phi(y_i^j); \boldsymbol{\alpha}_{kj})$$

Modèle génératif du mélange de copules gaussiennes





Échantillonnage de la classe

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

↓

Échantillonnage Gaussien $\boldsymbol{y}_i|\boldsymbol{z}_{ik} = 1$

↓

$$\boldsymbol{y}_i|\boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

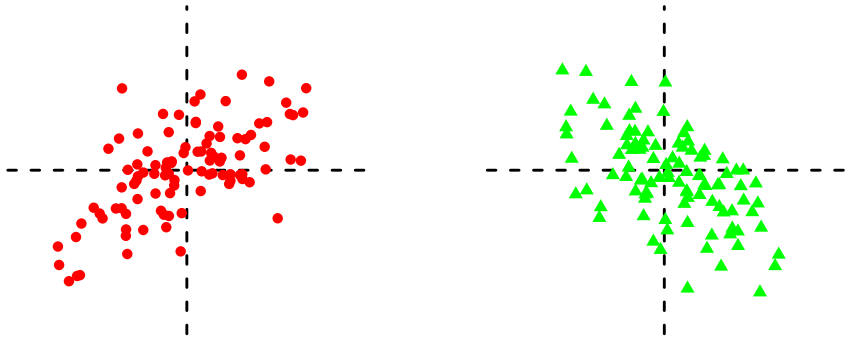
↓

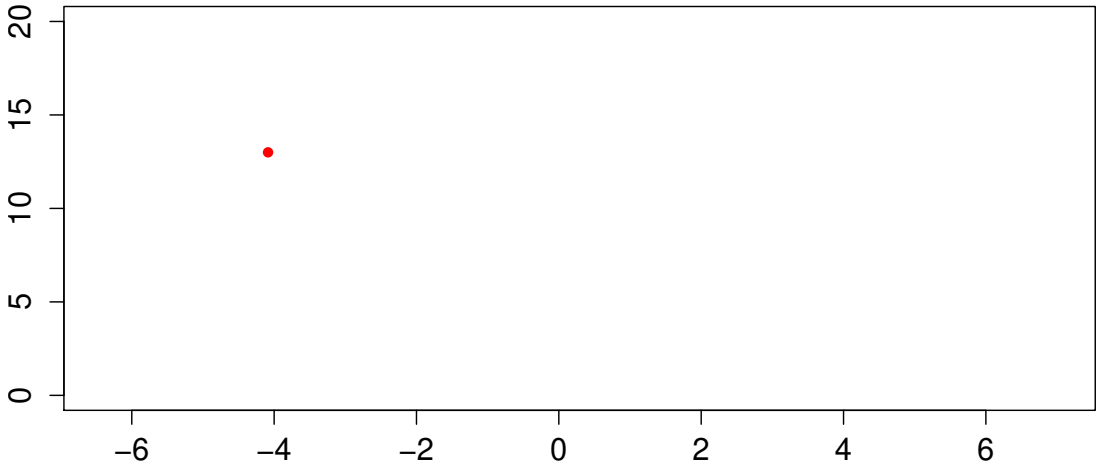
Calcul de l'observation $\boldsymbol{x}_i|(\boldsymbol{z}_i, \boldsymbol{y}_i)$

↓

$$x_i^j = F_{kj}^{-1}(\Phi(y_i^j); \boldsymbol{\alpha}_{kj})$$

Modèle génératif du mélange de copules gaussiennes





Échantillonnage de la classe

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

↓

Échantillonnage Gaussien $\boldsymbol{y}_i | \boldsymbol{z}_{ik} = 1$

↓

$$\boldsymbol{y}_i | \boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

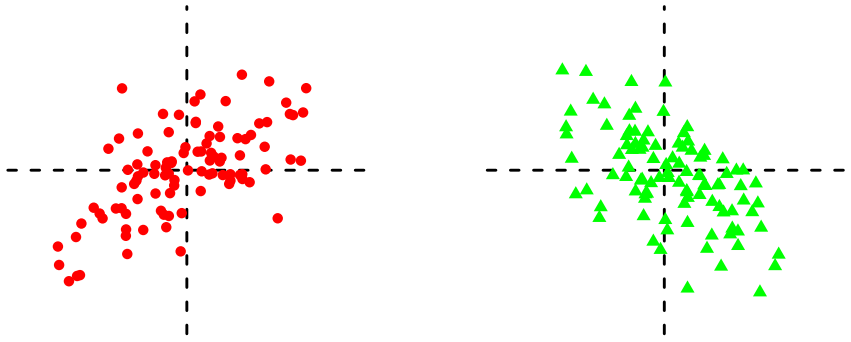
↓

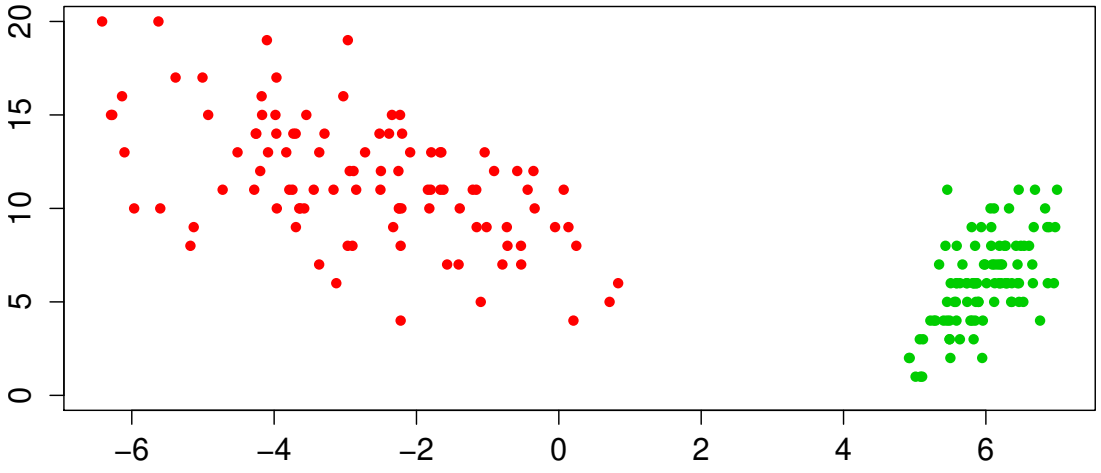
Calcul de l'observation $\boldsymbol{x}_i | (\boldsymbol{z}_i, \boldsymbol{y}_i)$

↓

$$x_i^j = F_{kj}^{-1}(\Phi(y_i^j); \boldsymbol{\alpha}_{kj})$$

Modèle génératif du mélange de copules gaussiennes





Échantillonnage de la classe

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

↓

Échantillonnage Gaussien $\boldsymbol{y}_i | \boldsymbol{z}_{ik} = 1$

↓

$$\boldsymbol{y}_i | \boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

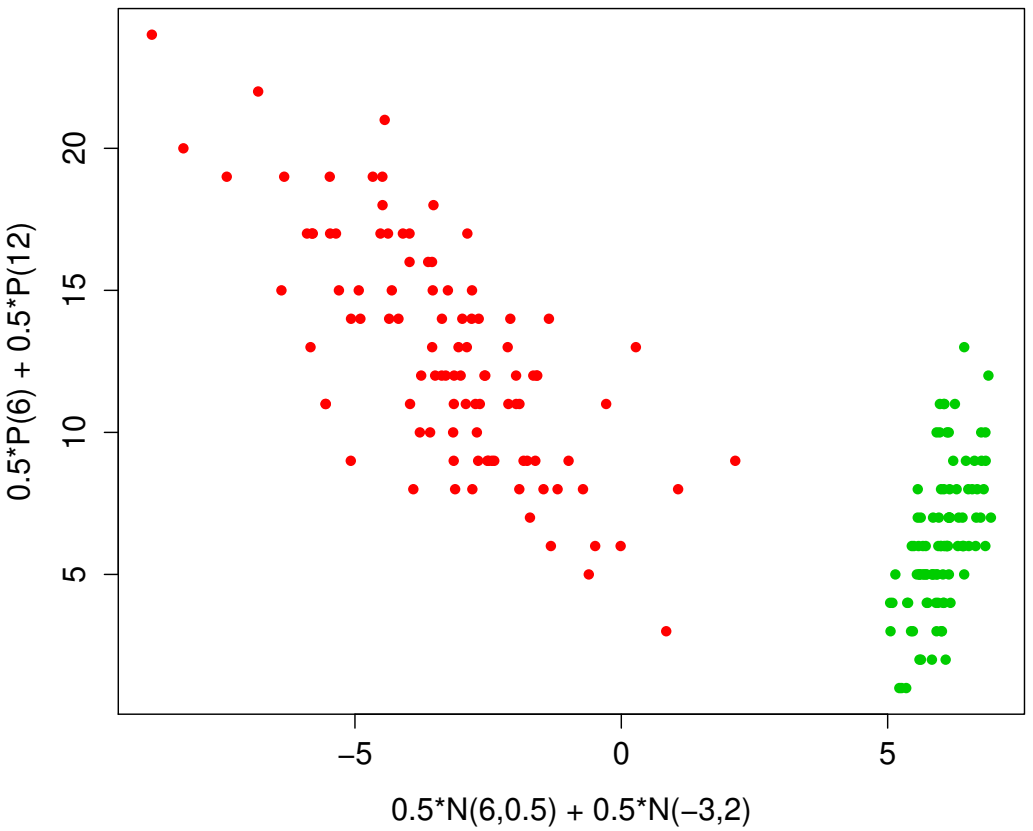
↓

Calcul de l'observation $\boldsymbol{x}_i | (\boldsymbol{z}_i, \boldsymbol{y}_i)$

↓

$$x_i^j = F_{kj}^{-1}(\Phi(y_i^j); \boldsymbol{\alpha}_{kj})$$

Modèle génératif du mélange de copules gaussiennes



Échantillonnage de la classe

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

↓

Échantillonnage Gaussien $\boldsymbol{y}_i | \boldsymbol{z}_{ik} = 1$

↓

$$\boldsymbol{y}_i | \boldsymbol{z}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$$

↓

Calcul de l'observation $\boldsymbol{x}_i | (\boldsymbol{z}_i, \boldsymbol{y}_i)$

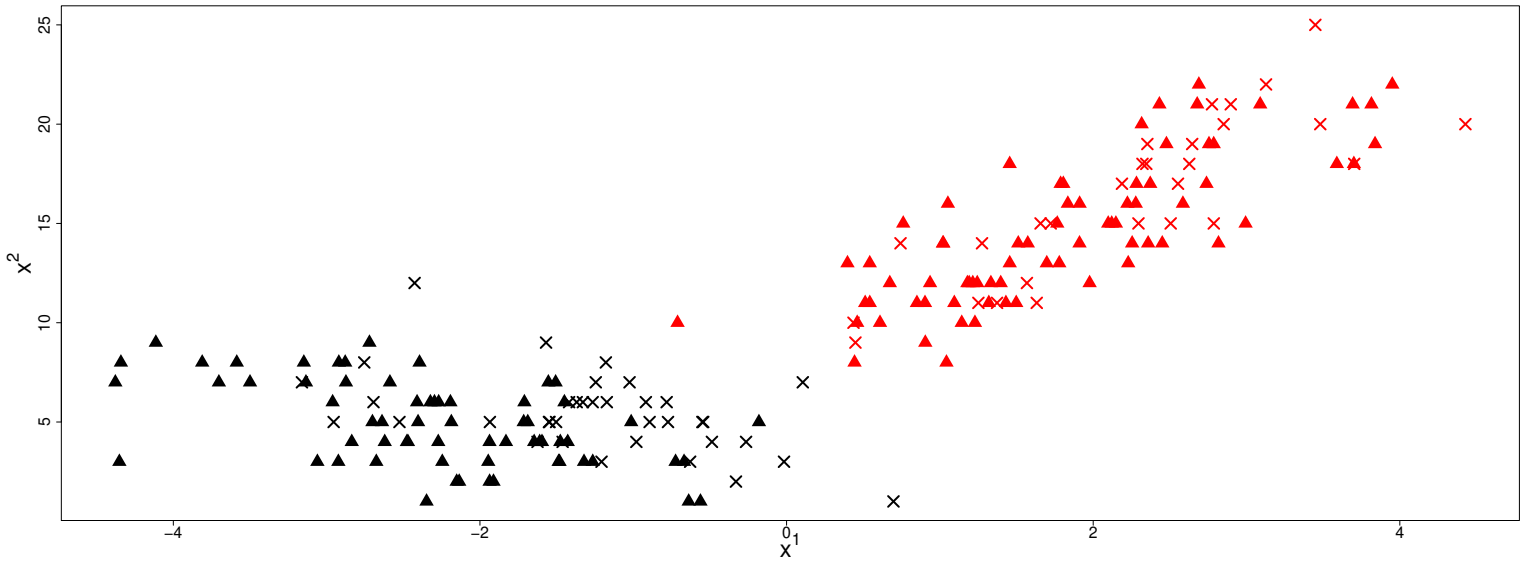
↓

$$x_i^j = F_{kj}^{-1}(\Phi(y_i^j); \boldsymbol{\alpha}_{kj})$$

Liens avec d'autres modèles

- Si $\mathbf{\Gamma}_k = \boldsymbol{I}$, $\forall k = 1, \ldots, g$: équivalence avec le modèle d'indépendance conditionnelle.
- Si uniquement des variables continues : équivalence avec le modèle de mélange gaussien.
- Si uniquement des variables ordinales : liens avec les modèle gaussiens par intervalles (Gouget).
- Si données continues et ordinales : liens avec le modèle gaussien latent (Everitt).

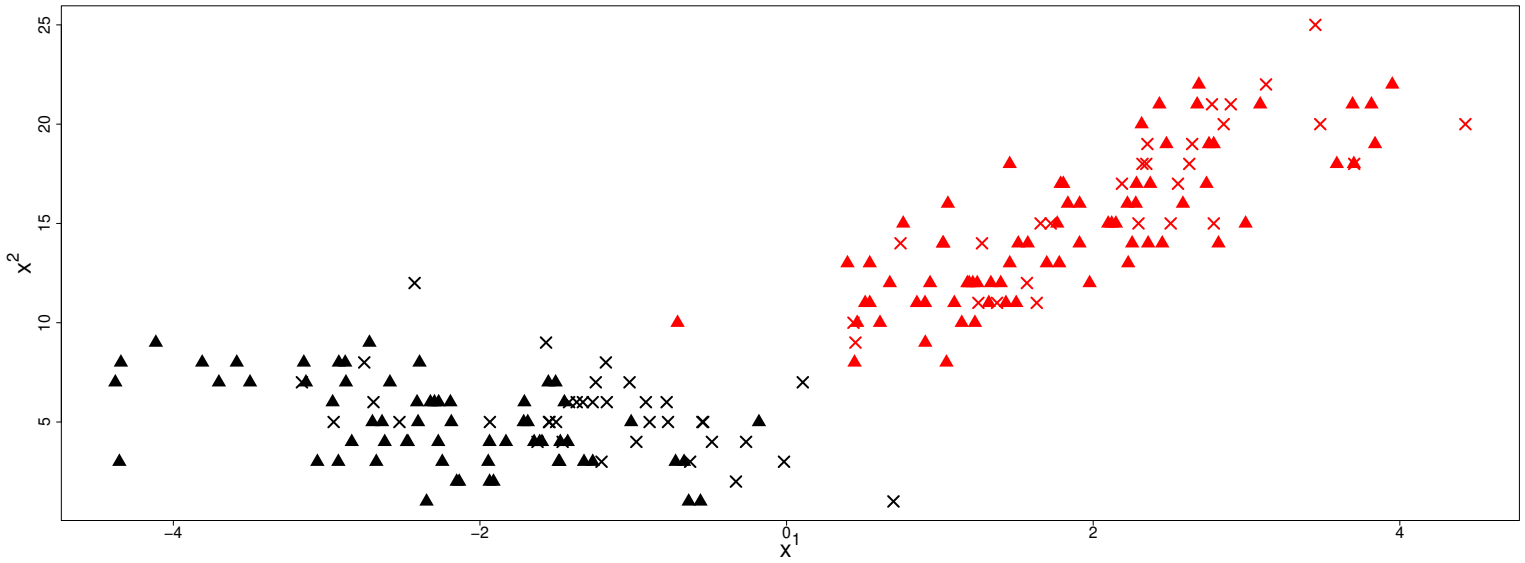
Visualisation d'une classe



Visualisation d'une classe

Calcul de l'espérance de $\boldsymbol{y}_i | z_{ik} = 1$

$$\mathbb{E}[\boldsymbol{y}_i | \boldsymbol{x}_i, z_{ik} = 1; \boldsymbol{\alpha}_k]$$



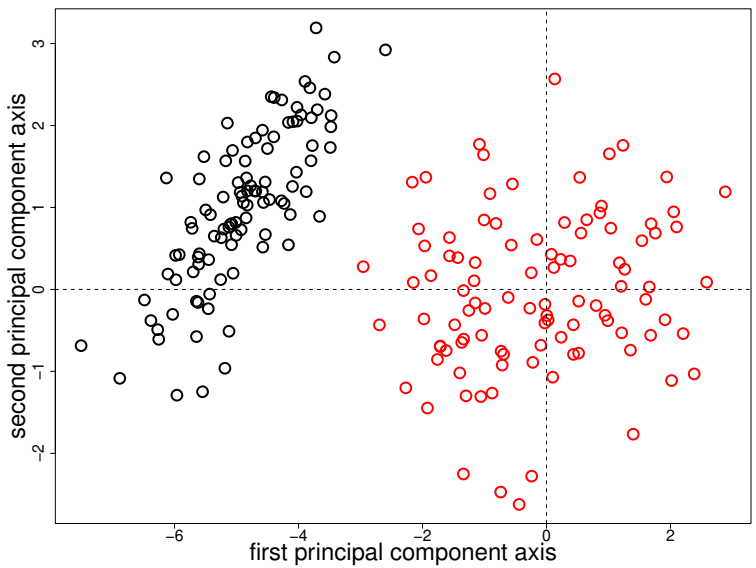
Visualisation d'une classe

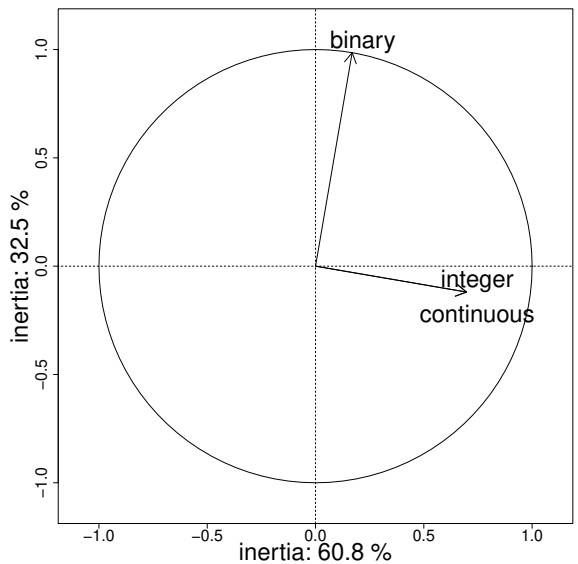
Calcul de l'espérance de $\boldsymbol{y}_i | z_{ik} = 1$

$$\mathbb{E}[\boldsymbol{y}_i | \boldsymbol{x}_i, z_{ik} = 1; \boldsymbol{\alpha}_k]$$

↓

Projections sur les axes factoriels déterminées par $\boldsymbol{\Gamma}_k$





Estimation par maximum de vraisemblance difficile

Maximisation de la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

- Difficile à cause du mélange.
- Solution : algorithm EM.

Maximisation de la log-vraisemblance complétée en z

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ik} \ln \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k).$$

- Étape E compliquée si $\boldsymbol{x}_i^{\mathrm{D}}$ grand car nécessite de calculer $f_k(\boldsymbol{x}_i^{\mathrm{D}} | \boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_k)$.
- Étape M compliquée à cause de la copule car pas de solution quand des variables sont discrètes.

Estimation Bayésienne

Échantillonnage selon la loi a posteriori

$$p(\boldsymbol{\theta}|\mathbf{x}).$$

- Indépendance entre les distributions des priors.
- Utilisation de priors conjugués pour avoir des distributions *a posteriori* explicites.

Échantillonneur de Gibbs

$$\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta} \tag{6}$$

$$\{\boldsymbol{\alpha}_{kj}\}, \mathbf{y}|\mathbf{x}, \mathbf{z}, \{\boldsymbol{\Gamma}_k\} \tag{7}$$

$$\boldsymbol{\pi}|\mathbf{z} \tag{8}$$

$$\{\boldsymbol{\Gamma}_k\}|\mathbf{y}, \mathbf{z} \tag{9}$$

Application : vins "Vinho Verde"

Vins "Vinho Verde"

Les données :

- 6497 vins (1599 rouges et 4898 blancs).
- 11 variables continues (caractéristiques chimiques).
- 1 variable entière (note de qualité).

Étude :
Clustering en cachant la nature du vin (rouge ou blanc).

	g	1	2	3	4	5	6
BIC	loc. indpt.	-63516	-61069	-61010	**-55967**	-60250	-57163
	hetero.	-44675	**-34520**	-39724	-44692	-44484	-48349
	homo.	-44675	-39372	**-38289**	-45209	-43217	-42417
ICL	loc. indpt.	-63516	-61229	-61365	**-56310**	-60726	-58138
	hetero.	-44675	**-34688**	-40176	-44933	-44758	-48959
	homo.	-44675	-39607	**-38791**	-45380	-43345	-42667

TABLE: Valeurs des critères d'information obtenues par les trois modèle pour l'étude `vins Vinho Verde`.

Vins "Vinho Verde"

	blanc	rouge
c1	**4359**	9
c2	538	**1590**

(a)

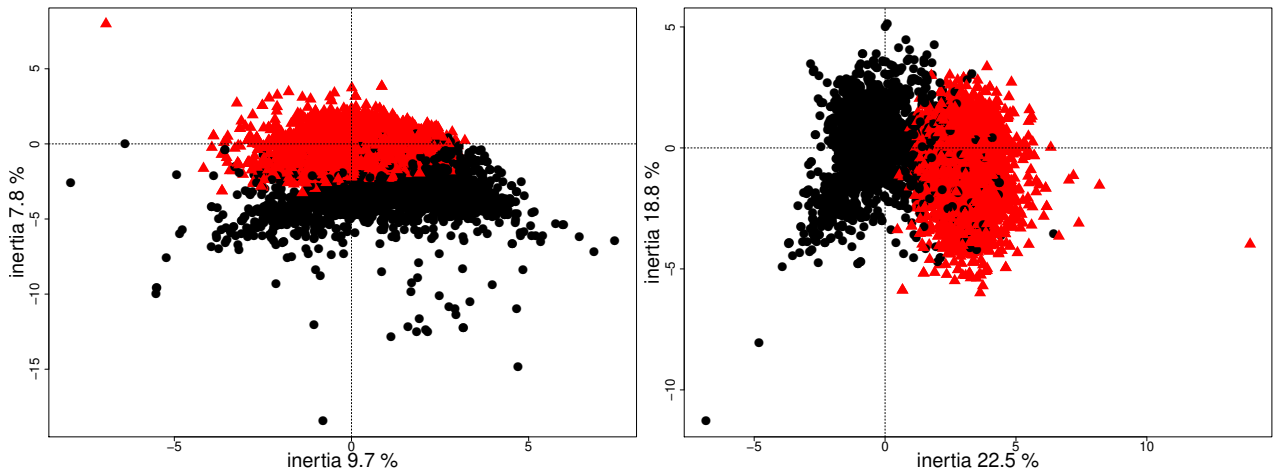
	blanc	rouge
c1	**2441**	12
c2	**1911**	7
c3	545	**1580**

(b)

	blanc	rouge
c1	**2547**	1561
c2	**2007**	35
c3	**275**	3
c4	**68**	0

(c)

TABLE: Matrices de confusion entre la nature des vins et la partition obtenues par : (a) hétéro. avec $g = 2$; (b) homo. avec $g = 3$; (c) indpt. cond. avec $g = 4$.

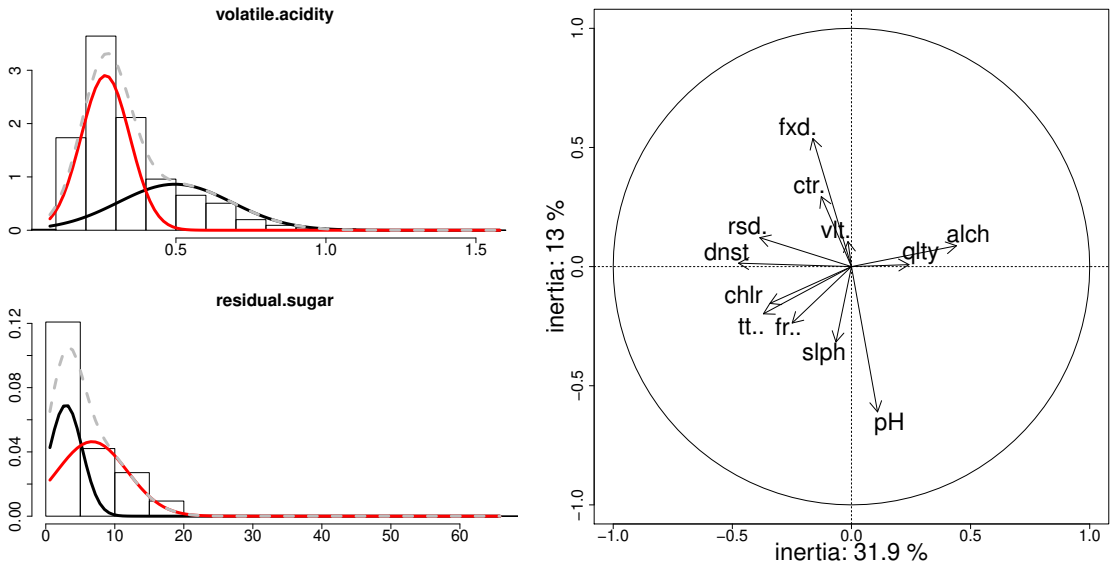


(a) Plan (3,5) de l'ACP de la classe 1 (b) Plan (1,2) de l'ACP de la classe 2

FIGURE: Projections dans les plans ACP de chacune des classes obtenues par le modèle hétéro. à deux classes.

Vins "Vinho Verde"

- Classe 1 (blancs) : vins peu acide, plus sucrés et plus fort.
- Classe 2 (rouges) : vins avec plus de chlorides et de sulfites.



(a) $\pi_k p(x_i^j | \boldsymbol{\alpha}_{kj}, z_{ik} = 1)$ (b) Cercle des corrélations pour $\boldsymbol{\Gamma}_2$

FIGURE: Résumé du modèle hétéro. avec $g = 2$. La classe 1 est en noir et la classe 2 en blanc.

Conclusion

- Mélange + Marginales classiques = Interprétation.
- Peu d'individus $\Rightarrow$ hypo. indépendance conditionnelle :
  - Marginales classiques.
  - rmixmod (http ://www.mixmod.org/).
- Bcp d'individus $\Rightarrow$ mélange de copules gaussiennes :
  - Marginales classiques.
  - Dépendances intra-classes.
  - Visualisation.
  - package R à venir.
  - préprint (http ://math.univ-lille1.fr/$\sim$marbaclo/).