

Sélection de variables pour la classification non supervisée des données mixtes

Université Paris-Saclay

18 mars 2022

Plan

1 Sélection de variables en clustering

- Clustering
- Contexte
- Notations
- La sélection

2 Deux critères d'information

- Optimisation de BIC directement
- Optimisation du critère ICL sans MLE

3 Expérimentations numériques

- Simulations
- Données réelles

Outlines

1 Sélection de variables en clustering

- Clustering
- Contexte
- Notations
- La sélection

2 Deux critères d'information

- Optimisation de BIC directement
- Optimisation du critère ICL sans MLE

3 Expérimentations numériques

- Simulations
- Données réelles

Clustering

Données accessibles : \mathbf{x}				Objectif : \mathbf{z}		
12.91	grand	3	...	0	1	0
15.00	petit	5	...	1	0	0
11.66	grand	4	...	0	0	1
12.50	grand	6	...	0	1	0
13.00	moyen	6	...	0	0	1
\vdots	\vdots	\vdots	\ddots	\vdots		

- $\mathbf{x} = (x_i; i = 1, \dots, n)$: les données observées.
- Chaque observation $x_i = (x_{i1}, \dots, x_{id})^\top \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d$,
- \mathcal{X}_j dépend de la nature de la variable, i.e $\mathcal{X}_j = \mathbb{R}, \mathbb{N}$ ou $\{1, \dots, m_j\}$.

Contexte

Idée principale:

Certaines variables ne sont pas discriminantes

Objectif:

Identifier les variables discriminantes (*i.e.* possédant une distribution variable d'une composante à une autre du mélange) + identifier les classes (les clusters).

Avantages:

- Améliorer l'estimation des paramètres.
- Interprétation simple.

Remarque:

Lorsqu'on observe un grand nombre de variables ($n < d$), l'information de dépendance intra-classes est en général difficile à observer. On se restreint aux modèles de mélange avec indépendance conditionnelle.

Notations

Données

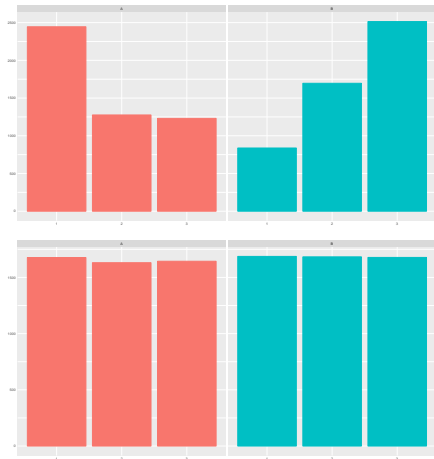
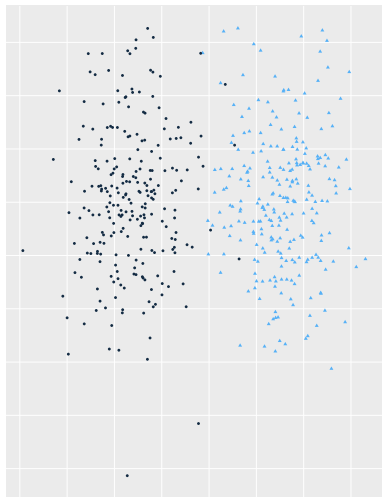
- $X = (X_1, \dots, X_d)^\top \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d$.
- $\mathbf{x} = (x_1, \dots, x_n)$ des réalisations iid de X .

La densité de probabilité

$$f(x_i|g, \theta) = \sum_{k=1}^g \tau_k f_k(x_i|\alpha_k) \text{ with } f_k(x_i|\alpha_k) = \prod_{j=1}^d f_{kj}(x_{ij}|\alpha_{kj}),$$

- $\theta = \{\tau_k, \alpha_k; k = 1, \dots, g\}$ regroupe les paramètres, τ_k est la taille de la composante k tel que $0 < \tau_k \leq 1$ et $\sum_{k=1}^g \tau_k = 1$.
- f_{kj} peut être une normale $\mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$, Poisson $\mathcal{P}(\alpha_{kj})$ ou multinomiale $\mathcal{M}(\alpha_{kj1}, \dots, \alpha_{kjm_j})$.

Variable discriminante vs bruit



Un modèle

Le rôle d'une variable

Un vecteur binaire $\omega = (\omega_j; j = 1, \dots, d)$ tel que

$$\omega_j = \begin{cases} 1 & \text{si } X_j \text{ est discriminante} \\ 0 & \text{sinon.} \end{cases} \quad \text{ou} \quad \omega_j = \begin{cases} 0 & \text{si } \alpha_{1j} = \dots = \alpha_{gj} \\ 1 & \text{sinon.} \end{cases}$$

Un modèle

Le rôle d'une variable

Un vecteur binaire $\omega = (\omega_j; j = 1, \dots, d)$ tel que

$$\omega_j = \begin{cases} 1 & \text{si } X_j \text{ est discriminante} \\ 0 & \text{sinon.} \end{cases} \quad \text{ou} \quad \omega_j = \begin{cases} 0 & \text{si } \alpha_{1j} = \dots = \alpha_{gj} \\ 1 & \text{sinon.} \end{cases}$$

Un modèle

Le couple $\mathbf{m} = (g, \omega)$ définit un modèle, *i.e.*, la vraisemblance est donnée par

$$f(x_i | \mathbf{m}, \theta) = \prod_{j \in \Omega^c} f_{1j}(x_{ij} | \alpha_{1j}) \sum_{k=1}^g \tau_k \prod_{j \in \Omega} f_{kj}(x_{ij} | \alpha_{kj}),$$

où $\Omega = \{j : \omega_j = 1\}$ et $\Omega^c = \{1, \dots, d\} \setminus \Omega$.

La sélection de modèles

Un grand nombre de modèles en compétition

$$\text{card}(\mathcal{M}) = g_{\max} 2^d \text{ si } \mathcal{M} = \{\mathbf{m} = (g, \boldsymbol{\omega}); 1 \leq g \leq g_{\max}, \boldsymbol{\omega} \in \{0, 1\}^d\}$$

Difficulté:

On cherche le modèle \mathbf{m}^* qui maximise un critère d'information (IC)

$$\mathbf{m}^* = \underset{\mathbf{m} \in \mathcal{M}}{\text{argmax}} \text{IC}(\mathbf{m}).$$

La sélection de modèles

Un grand nombre de modèles en compétition

$$\text{card}(\mathcal{M}) = g_{\max} 2^d \text{ si } \mathcal{M} = \{ \mathbf{m} = (g, \boldsymbol{\omega}); 1 \leq g \leq g_{\max}, \boldsymbol{\omega} \in \{0, 1\}^d \}$$

Difficulté:

On cherche le modèle \mathbf{m}^* qui maximise un critère d'information (IC)

$$\mathbf{m}^* = \underset{\mathbf{m} \in \mathcal{M}}{\text{argmax}} \text{IC}(\mathbf{m}).$$

Classiquement: IC est BIC ou ICL.

Approches exhaustive ou sous-optimales:

Méthodes pour sélectionner \mathbf{m}^* :

- **Exhaustive**, impossible lorsque d est grand $\text{card}(\mathcal{M}) = g_{\max} 2^d$.
- **Backward, forward** (Raftery, A. and Dean, D., 2006).

Optimisation en deux étapes

On vise \mathbf{m}^* tel que

$$\mathbf{m}^* = \arg \max_{\mathbf{m} \in \mathcal{M}} \text{IC}(\mathbf{m}).$$

Soit \mathcal{M}_g la restriction de \mathcal{M} au sous-ensemble de modèles avec g composantes. Ainsi, on définit \mathbf{m}_g^* par

$$\mathbf{m}_g^* = \arg \max_{\mathbf{m} \in \mathcal{M}_g} \text{IC}(\mathbf{m}).$$

À partir des \mathbf{m}_g^* pour $g = 1, \dots, g_{\max}$, on obtient \mathbf{m}^* tel que

$$\mathbf{m}^* = \arg \max_{g=1, \dots, g_{\max}} \text{IC}(\mathbf{m}_g^*).$$

Outlines

- 1 Sélection de variables en clustering
 - Clustering
 - Contexte
 - Notations
 - La sélection
- 2 Deux critères d'information
 - Optimisation de BIC directement
 - Optimisation du critère ICL sans MLE
- 3 Expérimentations numériques
 - Simulations
 - Données réelles

Bayesian Information Criterion

Objectif

En supposant une loi *a priori* uniforme sur \mathcal{M} , nous avons

$$p(\mathbf{m}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{m}),$$

où

$$p(\mathbf{x}|\mathbf{m}) = \int p(\mathbf{x}|\mathbf{m}, \theta) p(\theta|\mathbf{m}) d\theta.$$

- La vraisemblance intégrée $p(\mathbf{x}|\mathbf{m})$ est inaccessible.
- BIC est l'approximation de Laplace de $\ln p(\mathbf{x}|\mathbf{m})$

$$\text{BIC}(\mathbf{m}) = \ell(\hat{\theta}_{\mathbf{m}}; \mathbf{m}, \mathbf{x}) - \frac{\nu_{\mathbf{m}}}{2} \ln n,$$

où $\hat{\theta}_{\mathbf{m}}$ est le MLE et $\nu_{\mathbf{m}}$ est le nombre de paramètres du modèle \mathbf{m} .

Integrated Complete-data Likelihood (ICL)

Objectif

Prise en compte de l'objectif de clustering

Classiquement, on fait appel au critère ICL donné par

$$\text{ICL}(\mathbf{m}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}|\mathbf{m}) \text{ où } p(\mathbf{x}, \mathbf{z}|\mathbf{m}) = \int p(\mathbf{x}, \mathbf{z}|\mathbf{m}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{m})d\boldsymbol{\theta}.$$

- Une approximation

$$\text{ICL}(\mathbf{m}) \simeq \text{BIC}(\mathbf{m}) + \sum_{i=1}^n \sum_{k=1}^g \hat{z}_{ik} \ln t_{ik}(\hat{\boldsymbol{\theta}}_{\mathbf{m}}).$$

- $\hat{\mathbf{z}}$ est la partition donnée par la règle du Maximum A Posteriori.
- **Un EM pour calculer le MLE pour chaque modèle en compétition!!**

Optimiser BIC directement

Principe

L'hypothèse d'indépendance conditionnelle permet de maximiser directement le critère BIC.

Idée

On peut faire appel à une version modifiée de l'algorithme EM pour maximiser la log-vraisemblance pénalisée

$$\ell_{\text{pen}}(\boldsymbol{\theta} | \mathbf{m}, \mathbf{x}) = \ell(\boldsymbol{\theta} | \mathbf{m}, \mathbf{x}) - \nu_{\mathbf{m}} c,$$

pour toute constante c .

Green, P. J. (1990). On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452.

EM pour optimiser BIC

La log-vraisemblance pénalisée complétée

$$\ell_{\text{pen}}(\boldsymbol{\theta} | \mathbf{m}, \mathbf{x}, \mathbf{z}) = \ell(\boldsymbol{\theta} | \mathbf{m}, \mathbf{x}, \mathbf{z}) - (g - 1)c - c \sum_{j=1}^d \nu_j (g\omega_j + 1 - \omega_j),$$

où ν_j est le nombre de paramètres de la marginale univariée de la variable j .

- Cette version modifiée de l'algorithme EM maximise la log-vraisemblance pénalisée à g fixé.
- Un point initial aléatoire $(\mathbf{m}^{[0]}, \boldsymbol{\theta}^{[0]})$ où $\mathbf{m}^{[0]} = (g, \boldsymbol{\omega}^{[0]})$.

Expressions de $\ell(\boldsymbol{\theta}|\mathbf{m}, \mathbf{x})$ et $\ell(\boldsymbol{\theta}|\mathbf{m}, \mathbf{x}, \mathbf{z})$

$$\ell(\boldsymbol{\theta}|\mathbf{m}, \mathbf{x}) = \left(\sum_{j \in \Omega^c} \sum_{i=1}^n \ln f_{1j}(x_{ij}|\alpha_{1j}) \right) + \left(\sum_{i=1}^n \ln \left(\sum_{k=1}^g \tau_k \prod_{j \in \Omega} f_{kj}(x_{ij}|\alpha_{kj}) \right) \right).$$

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{m}, \mathbf{x}, \mathbf{z}) &= \sum_{j \in \Omega^c} \sum_{i=1}^n \ln f_{1j}(x_{ij}|\alpha_{1j}) + \sum_{k=1}^g \sum_{i=1}^n z_{ik} \ln \tau_k \\ &\quad + \sum_{j \in \Omega} \sum_{k=1}^g \sum_{i=1}^n z_{ik} \ln f_{kj}(x_{ij}|\alpha_{kj}) \end{aligned}$$

Étape E d'une itération [r]

Calcul de la partition floue

Étape E

$t_{ik}^{[r]} := \mathbb{E}[Z_{ik} | \mathbf{x}_i, \mathbf{m}, \boldsymbol{\theta}^{[r-1]}]$, donnée par

$$t_{ik}^{[r]} := \frac{\tau_k^{[r-1]} \prod_{j=1}^d f_{kj}(x_{ij} | \boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^g \tau_\ell^{[r-1]} \prod_{j=1}^d f_{\ell j}(x_{ij} | \boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

Étape M d'une itération [r]

Maximisation de l'espérance de la log-vraisemblance complétée sur (ω, θ)

Étape M

$$m^{[r]} = (g, \omega^{[r]}) \text{ où } \tau_k^{[r]} = \frac{\sum_{i=1}^n t_{ik}^{[r]}}{n},$$

$$\omega_j^{[r]} = \begin{cases} 1 & \text{si } \Delta_j^{[r]} > 0 \\ 0 & \text{sinon} \end{cases}, \text{ et } \alpha_{jk}^{[r]} = \begin{cases} \alpha_{kj}^{*[r]} & \text{si } \omega_j^{[r]} = 1 \\ \tilde{\alpha}_{kj} & \text{sinon} \end{cases},$$

$$\text{où } \Delta_j^{[r]} = \sum_{k=1}^g \sum_{i=1}^n t_{ik}^{[r]} (\ln f_{kj}(x_{ij} | \alpha_{kj}^{*[r]}) - \ln f_{1j}(x_{ij} | \tilde{\alpha}_{1j})) - (g-1)\nu_j c$$

- $\alpha_{jk}^{*[r]} = \arg \max_{\alpha_{kj}} \sum_{i=1}^n t_{ik}^{[r]} \ln f_{kj}(x_{ij} | \alpha_{kj})$ si j est discriminante
- $\tilde{\alpha}_{1j} = \arg \max_{\alpha_{1j}} \sum_{i=1}^n \ln f_{1j}(x_{ij} | \alpha_{1j})$ sinon.

Objectif : éviter de calculer le MLE

👉 Un autre nom

$$\text{MICL}(\mathbf{m}) = \ln p(\mathbf{x}, \mathbf{z}^* | \mathbf{m}) \text{ with } \mathbf{z}^* = \arg \max_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z} | \mathbf{m}).$$

$$p(\mathbf{x}, \mathbf{z} | \mathbf{m}) = p(\mathbf{z} | g) \prod_{j=1}^d p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}),$$

Plus précisément

$$p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}) = \int p(\boldsymbol{\alpha}_{\bullet j} | g, \omega_j) \prod_{k=1}^g \prod_{i=1}^n f_{kj}(x_{ij} | \alpha_{kj})^{z_{ik}} d\boldsymbol{\alpha}_{\bullet j}.$$

$p(\mathbf{z}|g)$ et $p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z})$

Expression de $p(\mathbf{z}|g)$

$$p(\mathbf{z}|g) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^g \Gamma(n_k + \frac{1}{2})}{\Gamma(n + \frac{g}{2})}, \text{ où } n_k = \sum_{i=1}^n z_{ik}.$$

Expression de $p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z})$ lorsque la variable j est continue

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \begin{cases} \pi^{-n/2} \left(\frac{b_j^{a_j/2} d_j^{1/2}}{\Gamma(a_j/2)} \right)^g \prod_{k=1}^g \frac{\Gamma(A_{kj}/2)}{B_{kj}^{A_{kj}} D_{kj}^{1/2}} & \text{if } \omega_j = 1 \\ \pi^{-n/2} \frac{b_j^{a_j} d_j^{1/2}}{\Gamma(a_j/2)} \frac{\Gamma(A_j/2)}{B_j^{A_j} D_j^{1/2}} & \text{if } \omega_j = 0 \end{cases},$$

$$\text{où } A_j = n + a_j, B_j^2 = b_j^2 + \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 + \frac{(c_j - \bar{x}_j)^2}{d_j^{-1} + n^{-1}}, D_j = n + d_j, \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij},$$

$$A_{kj} = n_k + a_j, B_{kj}^2 = b_j^2 + \sum_{i=1}^n z_{ik} (x_{ij} - \bar{x}_{jk})^2 + \frac{(c_j - \bar{x}_{jk})^2}{d_j^{-1} + n_k^{-1}}, D_{kj} = n_k + d_j, \bar{x}_{jk} = \frac{1}{n_k} \sum_{i=1}^n z_{ik} x_{ij}$$

$$\text{et } n_k = \sum_{i=1}^n z_{ik}.$$

Optimisation du critère MICL

Optimisation alternée pour trouver \mathbf{m}_g^* . Chaque itération $[r]$ consiste à alterner **deux** étapes:

Étape de partition (transfert):

On cherche $\mathbf{z}^{[r]}$ tel que

$$\ln p(\mathbf{x}, \mathbf{z}^{[r]} | \mathbf{m}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]} | \mathbf{m}^{[r]}).$$

Étape modèle (explicite):

On cherche $\mathbf{m}^{[r+1]} = \operatorname{argmax}_{\mathbf{m} \in \mathcal{M}_g} \ln p(\mathbf{x}, \mathbf{z}^{[r]} | \mathbf{m})$, où

$$\mathbf{m}^{[r+1]} = (g, \boldsymbol{\omega}^{[r+1]}) \text{ with } \omega_j^{[r+1]} = \operatorname{argmax}_{\omega_j \in \{0,1\}} p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}^{[r]}).$$

Plus de détails sur l'étape partition

À l'itération $[r]$: $\mathbf{z}^{[r]}$ est initialisé à $\mathbf{z}^{(0)} = \mathbf{z}^{[r-1]}$. On fait S itérations où une itération (s) est donnée par:

- **Tirage individuel:** $i^{(s)} \sim \mathcal{U}\{1, \dots, n\}$.
- **Optimisation de la partition:** conserver les affectations $\mathcal{Z}^{(s)} = \{\mathbf{z} : z_i = z_i^{(s-1)}, \forall i \neq i^{(s)}\}$ et affecter $\mathbf{z}^{(s)}$ tel que

$$\mathbf{z}^{(s)} = \operatorname{argmax}_{\mathbf{z} \in \mathcal{Z}^{(s)}} \ln p(\mathbf{x}, \mathbf{z} | \mathbf{m}^{[r]}).$$

On arrête si : $\ln p(\mathbf{x}, \mathbf{z} | \mathbf{m}^{[r]})$ n'augmente pas après S itérations

Plus de détails sur l'étape partition

À l'itération $[r]$: $\mathbf{z}^{[r]}$ est initialisé à $\mathbf{z}^{(0)} = \mathbf{z}^{[r-1]}$. On fait S itérations où une itération (s) est donnée par:

- **Tirage individuel:** $i^{(s)} \sim \mathcal{U}\{1, \dots, n\}$.
- **Optimisation de la partition:** conserver les affectations $\mathcal{Z}^{(s)} = \{\mathbf{z} : z_i = z_i^{(s-1)}, \forall i \neq i^{(s)}\}$ et affecter $\mathbf{z}^{(s)}$ tel que

$$\mathbf{z}^{(s)} = \operatorname{argmax}_{\mathbf{z} \in \mathcal{Z}^{(s)}} \ln p(\mathbf{x}, \mathbf{z} | \mathbf{m}^{[r]}).$$

On arrête si : $\ln p(\mathbf{x}, \mathbf{z} | \mathbf{m}^{[r]})$ n'augmente pas après S itérations

Bilan

- Deux étapes rapides et simples.
- À chaque itération, $\ln p(\mathbf{x}, \mathbf{z} | \mathbf{m})$ augmente.
- Nous avons besoin d'un grand nombre d'initialisations aléatoires.

Outlines

1 Sélection de variables en clustering

- Clustering
- Contexte
- Notations
- La sélection

2 Deux critères d'information

- Optimisation de BIC directement
- Optimisation du critère ICL sans MLE

3 Expérimentations numériques

- Simulations
- Données réelles

Hypothèse d'indépendance conditionnelle : cas gaussien

Un modèle à deux composantes gaussiennes à proportions égales. Sous la composante k , $r = 6$ variables suivant une gaussienne $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ où

$$\mu_{kj} = \begin{cases} (-1)^k \delta & \text{si } |j| \leq 6 \\ 0 & \text{sinon} \end{cases} \quad \text{et } \Sigma_k[j, j'] = \begin{cases} 1 & \text{if } j = j' \\ \rho & \text{if } |j - j'| = 1 \\ 0 & \text{sinon} \end{cases} .$$

Hypothèse d'indépendance conditionnelle : cas gaussien

ρ	d	BIC			VarSelLCM	ICL/MICL	
		no selection	clustvarsel forward	clustvarsel backward		no selection	VarSelLCM
0	10	0.78	0.53	0.71	0.78	0.78	0.78
	25	0.31	0.34	0.71	0.77	0.04	0.77
	50	0.00	0.13	0.04	0.80	0.00	0.80
	100	0.00	0.10	0.00	0.77	0.00	0.77
0.4	10	0.78	0.52	0.69	0.72	0.78	0.78
	25	0.78	0.44	0.63	0.77	0.78	0.78
	50	0.50	0.30	0.03	0.79	0.08	0.80
	100	0.00	0.18	0.00	0.73	0.00	0.77

ARI moyen sur 50 réplifications. δ est choisi pour un taux de mauvais classement de 5%.

Sélection de g et des variables

ρ	d	BIC						ICL/MICL			
		no selection	clustvarsel		VarSelLCM		no selection	VarSelLCM			
			g	forward g	backward rel.	g		rel.	g	g	rel.
0	10	2.00	2.45	0.52	2.55	0.66	2.00	0.60	2.00	2.00	0.60
	25	1.40	2.80	0.22	2.40	0.50	2.00	0.25	1.05	2.00	0.24
	50	1.00	2.90	0.08	2.95	0.43	2.00	0.13	1.00	2.00	0.12
	100	1.00	2.95	0.04	3.00	0.72	2.00	0.06	1.00	2.00	0.06
0.4	10	2.00	2.60	0.28	2.55	0.40	2.25	0.61	2.00	2.00	0.60
	25	2.00	2.85	0.17	2.65	0.40	2.05	0.25	2.00	2.00	0.24
	50	1.65	2.85	0.12	2.85	0.44	2.05	0.13	1.10	2.00	0.12
	100	1.00	2.95	0.04	3.00	0.70	2.15	0.06	1.00	2.00	0.06

0.60, 0.25, 0.13 et 0.06 sont les vraies proportions de variables pertinentes pour $d = 10, 25, 50$ et 100 .

ANR : Mode de vie humain, variation génétique et épigénétique

Données de l'étude

Étude sur les populations d'agriculteurs bantous (31 populations échantillonnées) et les chasseurs-cueilleurs (6 populations échantillonnées) de l'Afrique centrale. Les deux classes sont de tailles 1003 et 232 respectivement.

En tout $n = 1235$ et $d = 160470$ toutes catégorielles (SNPs) codées 0, 1 et 2.

ANR : Mode de vie humain, variation génétique et épigénétique

Alleles of SNP rs123456 (located on chromosome 1)

Paul: CTTAGATTCAT **G** TCACTAGCTAGG
CTTAGATTCAT **G** TCACTAGCTAGG

Jose: CTTAGATTCAT **G** TCACTAGCTAGG
CTTAGATTCAT **A** TCACTAGCTAGG

Julia: CTTAGATTCAT **G** TCACTAGCTAGG
CTTAGATTCAT **G** TCACTAGCTAGG

Roger: CTTAGATTCAT **C** TCACTAGCTAGG
CTTAGATTCAT **C** TCACTAGCTAGG

ANR : Mode de vie humain, variation génétique et épigénétique

Modèle sélectionné

$\hat{g} = 2$ avec 58954 SNP (ou variables).

```
> table(GabonzOPT, twopops)
```

	twopops	
GabonzOPT	0	1
0	36	1003
1	196	0

Références

- Dean, N. and Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11–35.
- Green, P. J. (1990). On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452.
- Marbac, M. and Sedki, M. (2016). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, pages 1–15.
- Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009b). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, 53:3872–3882.

Annexe: cas catégoriel et comptage

$$p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}) = \begin{cases} \frac{1}{\prod_{i=1}^n \Gamma(x_{ij} + 1)} \left(\frac{b_j^{a_j}}{\Gamma(a_j)} \right)^g \prod_{k=1}^g \Gamma(A_{kj}) B_{kj}^{-A_{kj}} & \text{si } \omega_j = 1 \\ \frac{1}{\prod_{i=1}^n \Gamma(x_{ij} + 1)} \frac{b_j^{a_j}}{\Gamma(a_j)} \Gamma(A_j) B_j^{-A_j} & \text{si } \omega_j = 0 \end{cases},$$

où $A_j = \sum_{i=1}^n x_{ij} + a_j$, $B_j = b_j^2 + n$, $A_{kj} = \sum_{i=1}^n z_{ik} x_{ij} + a_j$ et $B_j = b_j^2 + \sum_{i=1}^n z_{ik}$.

$$p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}) = \begin{cases} \left(\frac{\Gamma(m_j a)}{\Gamma(a)^{m_j}} \right)^g \prod_{k=1}^g \frac{\prod_{h=1}^{m_j} \Gamma(\sum_{i=1}^n z_{ik} \mathbf{1}_{\{x_{ij}=h\}} + a_j)}{\Gamma(\sum_{i=1}^n z_{ik} + m_j a_j)} & \text{si } \omega_j = 1 \\ \frac{\Gamma(m_j a)}{\Gamma(a)^{m_j}} \frac{\prod_{h=1}^{m_j} \Gamma(\sum_{i=1}^n \mathbf{1}_{\{x_{ij}=h\}} + a_j)}{\Gamma(n + m_j a_j)} & \text{si } \omega_j = 0 \end{cases}.$$

Annexe : lois *a priori* conjuguées

- $\boldsymbol{\tau} | \mathbf{m}$ suit une loi de Dirichlet $\mathcal{D}_g(u_1, \dots, u_g)$
- Cas continue : $p(\boldsymbol{\alpha}_{kj}) = p(\sigma_{kj}^2) p(\mu_{kj} | \sigma_{kj}^2)$ où σ_{kj}^2 suit une loi $\mathcal{IG}(a_j/2, b_j^2/2)$ et $\mu_{kj} | \mathbf{m}, \sigma_{kj}^2 \sim \mathcal{N}(c_j, \sigma_{kj}^2/d_j)$.
- Cas entier : $\boldsymbol{\alpha}_{kj}$ suit une loi gamma $\mathcal{G}a(a_j, b_j)$ où $\boldsymbol{\alpha}_{kj}$ suit une loi de Dirichlet $\mathcal{D}_{m_j}(a_j, \dots, a_j)$ si j est une variables catégorielle à m_j modalités.
- Jeffreys pour les proportions (*i.e.*, $u_k = 1/2$) et les hyper-paramètres des variables catégorielles (*i.e.*, $a_{jk} = 1/2$).
- Aucune règle pour les autres paramètres.
- $a_j = 1$, $b_j = 1$, $c_j = \text{mean}(\mathbf{x}_{\bullet j})$ et $d_j = 0.01$ et $a_j = b_j = 1$ for les hyper-paramètres de la loi de Poisson.