

# Méthodes probabilistes pour la classification non supervisée

`masedki.github.io`

`eespe`

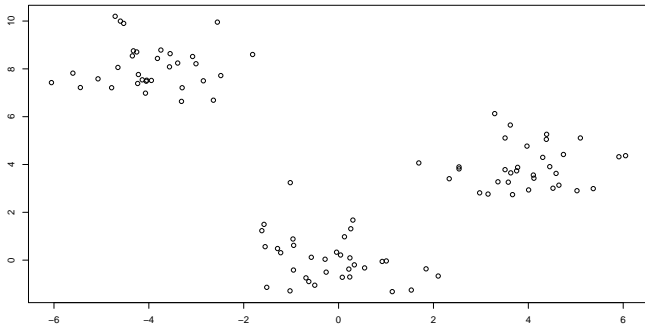
## Le problème

### Objectif:

- Estimation d'une partition  $\mathbf{z}$  parmi  $n$  observations  $\mathbf{x}$ .

### Notations:

- $g$ : nombre de groupes
- $\mathbf{x} = (\mathbf{x}_i; i = 1, \dots, n)$ : échantillon ( $n$  observations iid).
- $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ : vecteur des  $d$  variables décrivant l'observation  $i$ .
- $\mathbf{z} = (\mathbf{z}_i; i = 1, \dots, n)$ : partition (non observée).
- $\mathbf{z}_i = (z_{i1}, \dots, z_{ig})$  avec  $z_{ik} = 1$  indique de l'observation  $i$  appartient au groupe  $k$ .



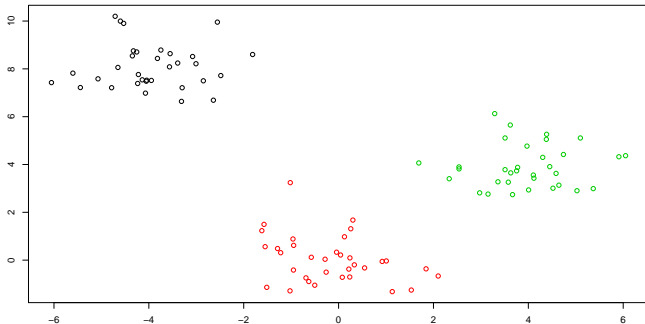
## Le problème

### Objectif:

- Estimation d'une partition  $\mathbf{z}$  parmi  $n$  observations  $\mathbf{x}$ .

### Notations:

- $g$ : nombre de groupes
- $\mathbf{x} = (\mathbf{x}_i; i = 1, \dots, n)$ : échantillon ( $n$  observations iid).
- $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ : vecteur des  $d$  variables décrivant l'observation  $i$ .
- $\mathbf{z} = (\mathbf{z}_i; i = 1, \dots, n)$ : partition (non observée).
- $\mathbf{z}_i = (z_{i1}, \dots, z_{ig})$  avec  $z_{ik} = 1$  indique de l'observation  $i$  appartient au groupe  $k$ .



# Le problème

## Les objectifs

- Estimation d'une règle de classement (*i.e.*, estimateur de  $z_i$  sachant  $x_i$ ).
- Evaluation du risque d'erreur de classement.
- Interprétation des groupes.
- Estimation du nombre de groupes  $g$ .
- Analyse de données complexes (mixed, valeurs manquantes ...).



## Modèle de mélange

### Idée principale:

modéliser la distribution des variables observées  $\mathbf{X}$ .

### Modèle génératif:

- $\mathbf{Z}_i \sim \mathcal{M}(\pi_1, \dots, \pi_g)$
- $\mathbf{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(\cdot)$ , e.g.,  $\mathcal{L}_k(\cdot) = \mathcal{N}(\mu_k, \Sigma_k)$ .

### Loi du couple:

La distribution du couple  $(\mathbf{X}, \mathbf{Z})$  est donnée par

$$f(\mathbf{x}_i, Z_{ik}) = \mathbb{P}(Z_{ik} = 1)f(\mathbf{x}_i | Z_{ik} = 1) = \pi_k f(\mathbf{x}_i; \theta_k)$$

où  $\theta_k$  représente les paramètres de la loi locale du groupe  $k$ .

### Loi marginale (le mélange):

$$f(\mathbf{x}_i; \theta) = \sum_{k=1}^g \mathbb{P}(Z_{ik} = 1)f(\mathbf{x}_i | Z_{ik} = 1) = \sum_{k=1}^g \pi_k f_k(\mathbf{x}_i; \theta_k)$$

## Modèle de mélange

### Idée principale:

modéliser la distribution des variables observées  $\mathbf{X}$ .

### Modèle génératif:

- $\mathbf{Z}_i \sim \mathcal{M}(\pi_1, \dots, \pi_g)$
- $\mathbf{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(\cdot)$ , e.g.,  $\mathcal{L}_k(\cdot) = \mathcal{N}(\mu_k, \Sigma_k)$ .

### Loi du couple:

La distribution du couple  $(\mathbf{X}, \mathbf{Z})$  est donnée par

$$f(\mathbf{x}_i, \mathbf{Z}_{ik}) = \mathbb{P}(Z_{ik} = 1)f(x_i | Z_{ik} = 1) = \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)$$

où  $\boldsymbol{\theta}_k$  représente les paramètres de la loi locale du groupe  $k$ .

### Loi marginale (le mélange):

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^g \mathbb{P}(Z_{ik} = 1)f(x_i | Z_{ik} = 1) = \sum_{k=1}^g \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)$$

## Modèle de mélange

### Idée principale:

modéliser la distribution des variables observées  $\mathbf{X}$ .

### Modèle génératif:

- $\mathbf{Z}_i \sim \mathcal{M}(\pi_1, \dots, \pi_g)$
- $\mathbf{X}_i | Z_{ik} = 1 \sim \mathcal{L}_k(\cdot)$ , e.g.,  $\mathcal{L}_k(\cdot) = \mathcal{N}(\mu_k, \Sigma_k)$ .

### Loi du couple:

La distribution du couple  $(\mathbf{X}, \mathbf{Z})$  est donnée par

$$f(\mathbf{x}_i, \mathbf{Z}_{ik}) = \mathbb{P}(Z_{ik} = 1)f(x_i | Z_{ik} = 1) = \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)$$

où  $\boldsymbol{\theta}_k$  représente les paramètres de la loi locale du groupe  $k$ .

### Loi marginale (le mélange):

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^g \mathbb{P}(Z_{ik} = 1)f(x_i | Z_{ik} = 1) = \sum_{k=1}^g \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)$$



## Deux types de partitions (maximum a posteriori)

**Fuzzy and hard partitions:**

$$\mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\sum_{\ell=1}^g \pi_\ell f_\ell(\mathbf{x}_i; \boldsymbol{\theta}_\ell)}$$

La groupe de l'observation  $\mathbf{x}_i$  est alors defini par

$$\hat{z}_{ik^*} = 1 \text{ if } k^* = \arg \max_k \mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i)$$

Deux niveaux d'approximation:

$$\mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}{\sum_{\ell=1}^g \pi_\ell f_\ell(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_\ell)}$$

## Deux types de partitions (maximum a posteriori)

**Fuzzy and hard partitions:**

$$\mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\sum_{\ell=1}^g \pi_\ell f_\ell(\mathbf{x}_i; \boldsymbol{\theta}_\ell)}$$

La groupe de l'observation  $\mathbf{x}_i$  est alors defini par

$$\hat{z}_{ik^*} = 1 \text{ if } k^* = \arg \max_k \mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i)$$

**Deux niveaux d'approximation:**

$$\mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}{\sum_{\ell=1}^g \pi_\ell f_\ell(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_\ell)}$$

## Maximum de vraisemblance

Pour l'échantillon  $\mathbf{x} = (\mathbf{x}_i; i = 1, \dots, n)$ , on souhaite obtenir  $\hat{\boldsymbol{\theta}} = \operatorname{argmax} \ell(\boldsymbol{\theta}; \mathbf{x})$  où

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^n \ln \left( \sum_{k=1}^g \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right).$$

On considère la log-vraisemblance complétée

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

### Algorithme EM

- Introduit dans le cas de données manquantes (pour nous c'est  $\mathbf{z}$  qui manque).
- Itérative.
- La log-vraisemblance augmente à chaque itération.
- À l'itération  $[r]$ , deux étapes:

- E-step: calcul de

$$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \mathbf{x}_i, \boldsymbol{\theta}^{[r-1]}).$$

- M-step:  $\boldsymbol{\theta}^{[r]}$  maximise la log-vraisemblance complétée

$$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^g t_{ik}^{[r-1]} \ln \left( \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right).$$

## Maximum de vraisemblance

Pour l'échantillon  $\mathbf{x} = (\mathbf{x}_i; i = 1, \dots, n)$ , on souhaite obtenir  $\hat{\boldsymbol{\theta}} = \operatorname{argmax} \ell(\boldsymbol{\theta}; \mathbf{x})$  où

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^n \ln \left( \sum_{k=1}^g \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right).$$

On considère la log-vraisemblance complétée

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

### Algorithme EM

- Introduit dans le cas de données manquantes (pour nous c'est  $\mathbf{z}$  qui manque).
- Itérative.
- La log-vraisemblance augmente à chaque itération.
- À l'itération  $[r]$ , deux étapes:

- E-step: calcul de

$$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \mathbf{x}_i, \boldsymbol{\theta}^{[r-1]}).$$

- M-step:  $\boldsymbol{\theta}^{[r]}$  maximise la log-vraisemblance complétée

$$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^g t_{ik}^{[r-1]} \ln \left( \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \right).$$

## Algorithme EM (commentaires)

- Beaucoup d'efforts sur l'initialisation
- M-step: maximise la log-vraisemblance complétée

$$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^g t_{ik}^{[r-1]} \ln (\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)).$$

nous permet d'obtenir les paramètres  $\hat{\boldsymbol{\theta}}^{[r]} = \{\hat{\pi}_1^{[r]}, \dots, \hat{\pi}_K^{[r]}, \hat{\boldsymbol{\theta}}_1^{[r]}, \dots, \hat{\boldsymbol{\theta}}_K^{[r]}\}$ .

- Soit une suite d'itérations EM :  $\hat{\boldsymbol{\theta}}^{[0]}, \hat{\boldsymbol{\theta}}^{[1]}, \dots, \hat{\boldsymbol{\theta}}^{[h]}, \dots$ , on montre que

$$\ell(\hat{\boldsymbol{\theta}}^{[h+1]}, \mathbf{x}) \geq \ell(\hat{\boldsymbol{\theta}}^{[h]}, \mathbf{x}), \quad \text{pour toute itération } h.$$

## Sélection de modèle

### Modèle

Un modèle est défini par un nombre de composantes, la distribution des composantes, les contraintes entre paramètres...

### Question

Comment effectuer la sélection de modèle?

### Approche standard

Définir l'ensemble des modèles en compétition  $\mathcal{M}$  en fixant un nombre de composantes maximal  $g_{\max}$ . Le modèle sélectionné maximisera un critère d'information.

### Approche exhaustive

Calcule d'un critère d'information pour chaque modèle dans  $\mathcal{M}$ .

### Outils

Critère d'information (BIC, ICL,...) qui pénalisent la log-vraisemblance.

## Sélection de modèle

### Modèle

Un modèle est défini par un nombre de composantes, la distribution des composantes, les contraintes entre paramètres...

### Question

Comment effectuer la sélection de modèle?

### Approche standard

Définir l'ensemble des modèles en compétition  $\mathcal{M}$  en fixant un nombre de composantes maximal  $g_{\max}$ . Le modèle sélectionné maximisera un critère d'information.

### Approche exhaustive

Calcule d'un critère d'information pour chaque modèle dans  $\mathcal{M}$ .

### Outils

Critère d'information (BIC, ICL,...) qui pénalisent la log-vraisemblance.

## Bayesian Information Criterion

Avec une loi a priori uniforme  $\omega \in \mathcal{M}$ :

$$p(\omega|\mathbf{x}) \propto p(\mathbf{x}|\omega) \text{ où } p(\mathbf{x}|\omega) = \int p(\mathbf{x}|\omega, \theta)p(\theta|\omega)d\theta.$$

Pour estimer  $\ln p(\mathbf{x}|\omega)$ , on utilise une approximation de Laplace. Cela donne lui au critère BIC:

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

où  $\hat{\theta}_\omega$  est l'EMV et où  $\nu_\omega$  est le nombre de paramètres du modèle  $\omega$ .

### Résumé

- Compromis: Précision/Complexité.
- Critère consistant.
- Besoin de l'EMV.
- L'objectif du clustering n'est pas modélisé dans BIC.



## Vraisemblance complétée intégrée

Le critère ICL est défini par

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

où  $\hat{\mathbf{z}}$  est la partition obtenue en considérant la règle du MAP et l'EMV  $\hat{\theta}_\omega$ .  
 $p(\mathbf{x}, \mathbf{z}|\omega)$  a une forme explicite si les composantes font parties de la famille exponentielle (+ priors conjugués).

Sinon, on utilise une approximation de Laplace

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

### Résumé

- Compromis: Précision/Complexité/Chevauchement des groupes
- On considère l'objectif de clustering.
- Besoin de l'EMV.
- Critère non consistant
- Robustesse à l'erreur de modèle.

## Sélection de variables en clustering

### Quelles variables doit on utiliser en clustering?

- Problème bien posé en classification supervisée car on dispose d'un critère objectif (taux d'erreur, ROC,...) ...
- Problème mal posé en clustering puisque la variable de classe n'est pas connue en avance. Ainsi, il faut déterminer les variables discriminantes au sens d'une variable non observée?
- Solution pragmatique 1: sélection a priori faite par l'utilisateur.
- Solution pragmatique 2: analyse a posteriori de la corrélation entre les variables et le variable de classe (estimée à partir de toutes les variables).

**Il faut estimer simultanément la partition et le rôle des variables**

## Sélection de variables en clustering

### Idée de départ

Seulement un sous-ensemble des variables explique la partition non observée.

### Avantages

- Améliore la précision de l'étude en réduisant la variance des estimateurs.
- Facilite l'interprétation des groupes.

