Méthodes probabilistes pour la classification non supervisée

Mohammed SEDKI

SDDP-MSP

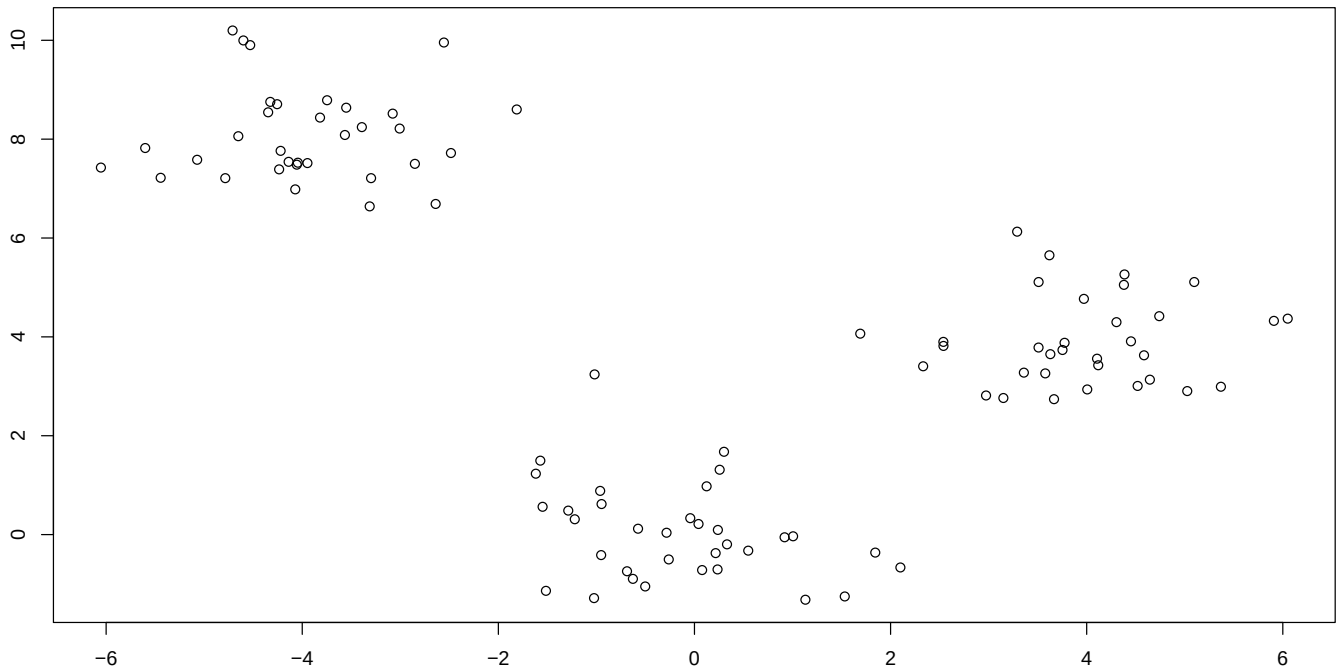
Le problème

Objectif:

- Estimation d'une partition **z** parmi n observations **x**.

Notations:

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



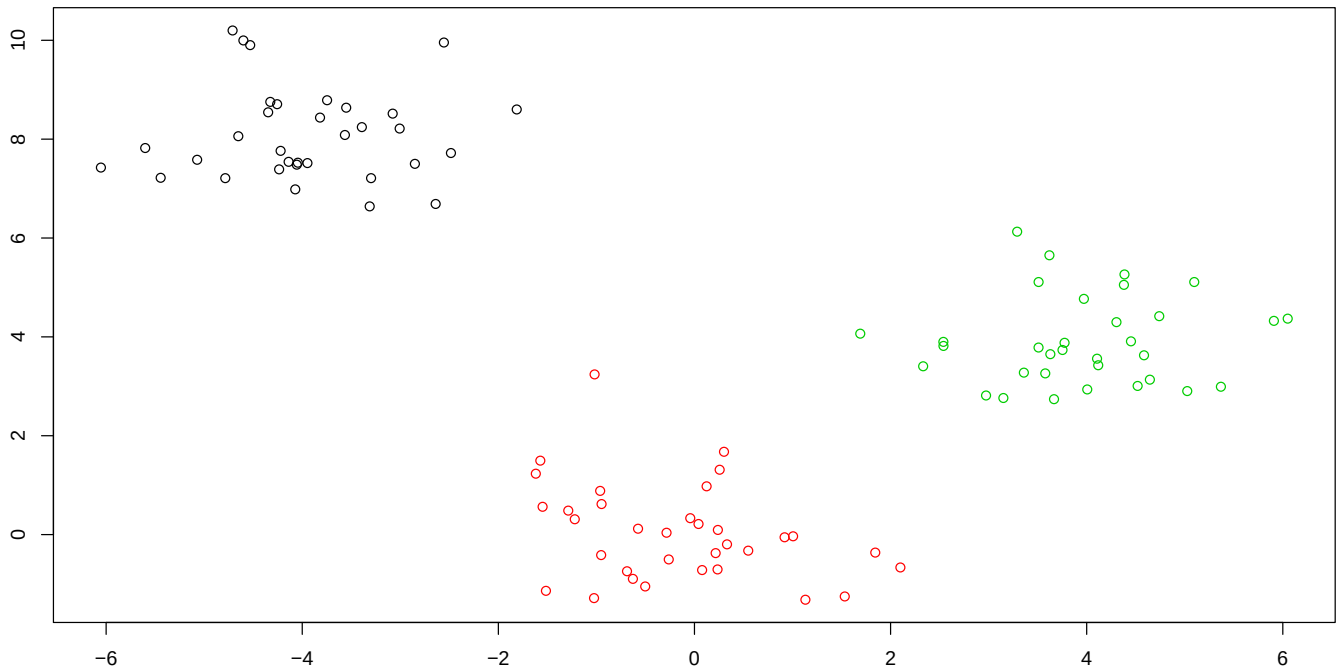
Le problème

Objectif:

- Estimation d'une partition **z** parmi n observations **x**.

Notations:

- g: nombre de groupes
- $\mathbf{x} = (\boldsymbol{x}_i; i = 1, \ldots, n)$: échantillon ($n$ observations iid).
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$: vecteur des d variables décrivant l'observation i.
- $\mathbf{z} = (\boldsymbol{z}_i; i = 1, \ldots, n)$: partition (non observée).
- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ 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 $\boldsymbol{z}_i$ sachant $\boldsymbol{x}_i$).
- 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 $\boldsymbol{X}$.

Modèle génératif:

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

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

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

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

Loi marginale (le mélange):

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

Modèle de mélange

Idée principale:
modéliser la distribution des variables observées $\boldsymbol{X}$.

Modèle génératif:

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

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

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

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

Loi marginale (le mélange):

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

Modèle de mélange

Idée principale:
modéliser la distribution des variables observées $\boldsymbol{X}$.

Modèle génératif:

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

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

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

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

Loi marginale (le mélange):

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

Deux types de partitions (maximum a posteriori)

Fuzzy and hard partitions:

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

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

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

Deux niveaux d'approximation:

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

Deux types de partitions (maximum a posteriori)

Fuzzy and hard partitions:

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

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

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

Deux niveaux d'approximation:

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

Maximum de vraisemblance

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

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

On considère la log-vraisemblance complétée

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

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: calcule de
  $$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]}).$$
  - M-step: $\boldsymbol{\theta}^{[r]}$ maximise la log-vraisemblance complétée
  $$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} t_{ik}^{[r-1]} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \right).$$

Maximum de vraisemblance

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

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

On considère la log-vraisemblance complétée

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

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: calcule de
  $$t_{ik}^{[r-1]} := \mathbb{P}(Z_{ik} = 1 | \boldsymbol{x}_i, \boldsymbol{\theta}^{[r-1]}).$$
  - M-step: $\boldsymbol{\theta}^{[r]}$ maximise la log-vraisemblance complétée
  $$\ln p(\mathbf{x}, \mathbf{t}^{[r-1]} | \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} t_{ik}^{[r-1]} \ln \left(\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \right).$$

Algorithme EM (commentaires)

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

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

  nous permet d'obtenir les paramètres $\widehat{\boldsymbol{\theta}}^{[r]} = \left\{\widehat{\pi}_1^{[r]}, \ldots, \widehat{\pi}_K^{[r]}, \widehat{\boldsymbol{\theta}}_1^{[r]}, \ldots, \widehat{\boldsymbol{\theta}}_K^{[r]}\right\}$.

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

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

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 $\boldsymbol{\omega} \in \mathcal{M}$:

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

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

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

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

Résumé

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

Vraisemblance complétée intégrée

Le critère ICL est défini par

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

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

Sinon, on utilise une approximation de Laplace

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

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 posterior 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.

Advantages

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