HDclassif : un package R pour la classification non-supervisée de données en grande dimension

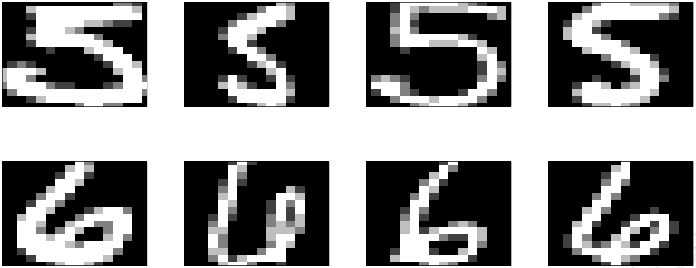
Laurent Bergé

MAP5 (UMR 8145), Université Paris-Descartes & Sorbonne Paris Cité

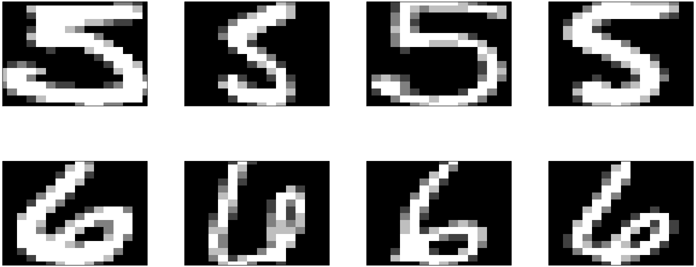
GREThA (UMR 5113), University of Bordeaux

10 Juin 2016, Journée R, Muséum National D'Histoire Naturelle, Paris

Qu’est-ce que la classification non-supervisée ?



Qu’est-ce que la classification non-supervisée ?



But de la classification non-supervisée

1. Catégoriser des observations dans des groupes cohérents
2. (optionnel) Avoir le “bon” nombre de groupes

Pourquoi faire de la classification non-supervisée ?

- Synthétiser des données complexes et volumineuses
- Partitionner les individus en des classes homogènes et interprétables

Schématisation du problème

Classification non-supervisée

- input: n observations $x_i \in \mathbb{R}^p$
- output: la partition $\{z_1, \ldots z_n\}$ (et le nombre de classes K)

Plan

1. Modèle de Mélange Gaussien
2. Modèles HDDC
3. Package HDclassif

Modèle de Mélange Gaussien

Modèle génératif:

Chaque observation i est générée de façon *indépendante* de la façon suivante :

1. $z_i \sim \mathcal{M}(\pi_1, \ldots, \pi_K)$
2. Sachant $z_i = k$:

$$x_i \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Paramètres du modèle

Pour chaque classe $k \in \{1, \ldots, K\}$:

- π_k proportion
- μ_k vecteur moyenne
- Σ_k matrice de variance-covariance

Modèle de Mélange Gaussien

Modèle génératif:

Chaque observation i est générée de façon *indépendante* de la façon suivante :

1. $z_i \sim \mathcal{M}\left(\pi_1, \ldots, \pi_K\right)$
2. Sachant $z_i = k$:

$$x_i \sim \mathcal{N}\left(\mu_k, \Sigma_k\right)$$

Paramètres du modèle

Pour chaque classe $k \in \{1, \ldots, K\}$:

- π_k proportion
- μ_k vecteur moyenne
- Σ_k matrice de variance-covariance

Modèle de Mélange Gaussien

Modèle génératif:

Chaque observation i est générée de façon *indépendante* de la façon suivante :

1. $z_i \sim \mathcal{M}(\pi_1, \ldots, \pi_K)$
2. Sachant $z_i = k$:

$$x_i \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Paramètres du modèle

Pour chaque classe $k \in \{1, \ldots, K\}$:

- π_k proportion
- μ_k vecteur moyenne
- Σ_k matrice de variance-covariance

Modèle de Mélange Gaussien

Modèle génératif:

Chaque observation i est générée de façon *indépendante* de la façon suivante :

1. $z_i \sim \mathcal{M}(\pi_1, \ldots, \pi_K)$
2. Sachant $z_i = k$:

$$x_i \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Paramètres du modèle

Pour chaque classe $k \in \{1, \ldots, K\}$:

- π_k proportion
- μ_k vecteur moyenne
- Σ_k matrice de variance-covariance

Estimateurs

Probabilité d'observer x_i sachant $z_i = k$:

$$
\begin{aligned}
& P\left(x_i | z_i = k, \mu_k, \Sigma_k\right) \\
& = f\left(x_i | \mu_k, \Sigma_k\right) \\
& = \exp\left((x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) + \log(\det \Sigma_k) + C^{te}\right)
\end{aligned}
$$

Estimateurs

Les estimateurs $\hat{\mu}_k$ et $\hat{\Sigma}_k$ sont les paramètres qui maximisent la vraisemblance.

Règle de Bayes

Si on ne connait pas z_i:

$$P(z_i = k | x_i, \mu, \Sigma) = \frac{\pi_k f(x_i | \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} f(x_i | \mu_{k'}, \Sigma_{k'})} \equiv t_{ik}$$

Par définition: $\sum_k t_{ik} = 1$.

Algorithme EM

1. Initialisation des z_i
2. Calcul des paramètres (π_k, μ_k et Σ_k) sachant z_i
3. Boucler jusqu'à convergence:
    1. Calcul des t_{ik}
    2. Estimation des paramètres (π_k, μ_k et Σ_k) sachant t_{ik}

Algorithme EM

1. Initialisation des z_i
2. Calcul des paramètres (π_k, μ_k et Σ_k) sachant z_i
3. Boucler jusqu'à convergence:
    1. Calcul des t_{ik}
    2. Estimation des paramètres (π_k, μ_k et Σ_k) sachant t_{ik}

Algorithme EM

1. Initialisation des z_i
2. Calcul des paramètres (π_k, μ_k et Σ_k) sachant z_i
3. Boucler jusqu'à convergence:
    1. Calcul des t_{ik}
    2. Estimation des paramètres (π_k, μ_k et Σ_k) sachant t_{ik}

Algorithme EM

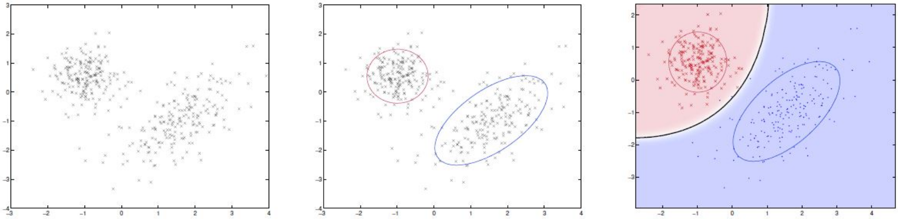
1. Initialisation des z_i
2. Calcul des paramètres (π_k, μ_k et Σ_k) sachant z_i
3. Boucler jusqu'à convergence:
   1. Calcul des t_{ik}
   2. Estimation des paramètres (π_k, μ_k et Σ_k) sachant t_{ik}

Algorithme EM

1. Initialisation des z_i
2. Calcul des paramètres (π_k, μ_k et Σ_k) sachant z_i
3. Boucler jusqu'à convergence:
    1. Calcul des t_{ik}
    2. Estimation des paramètres (π_k, μ_k et Σ_k) sachant t_{ik}

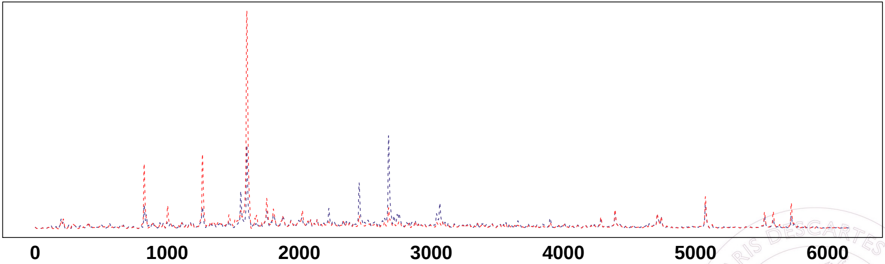
Illustration Modèle de Mélange Gaussien

Classification non-supervisée :



Difficultés du monde réel

- Les données réelles sont souvent en grande dimension (p est très grand)
- Le nombre d'observation est souvent faible ($n \ll p$)

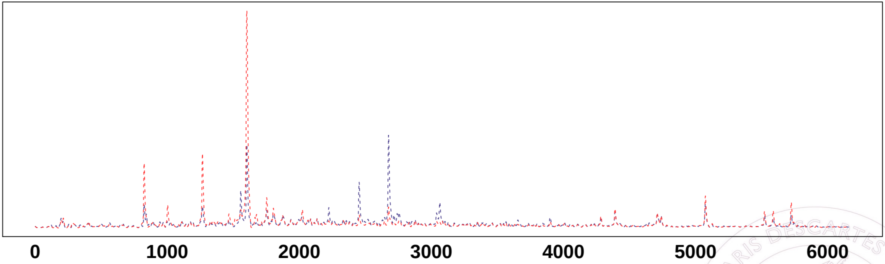


Problèmes

- Quand p est grand $\Rightarrow$ nombre de paramètres explose, complexité p^2
- Quand $n < p \Rightarrow$ modèle pas estimable (dû à l'estimation de Σ_k)

Difficultés du monde réel

- Les données réelles sont souvent en grande dimension (p est très grand)
- Le nombre d'observation est souvent faible ($n \ll p$)



Problèmes

- Quand p est grand $\Rightarrow$ nombre de paramètres explose, complexité p^2
- Quand $n < p \Rightarrow$ modèle pas estimable (dû à l'estimation de Σ_k)

Plan

Décomposition spectrale

- Décomposition spectrale de Σ_k:

$$\Sigma_k = Q_k \Delta_k Q_k^t,$$

avec:

- Q_k la matrice des vecteurs propres de Σ_k et $Q_k^{-1} = Q_k^t$
- Δ_k la matrice diagonale des valeurs propres de Σ_k:

$$\Delta_k = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Décomposition spectrale

- Décomposition spectrale de Σ_k:

$$\Sigma_k = Q_k \Delta_k Q_k^t,$$

avec:

- Q_k la matrice des vecteurs propres de Σ_k et $Q_k^{-1} = Q_k^t$
- Δ_k la matrice diagonale des valeurs propres de Σ_k:

$$\Delta_k = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Décomposition spectrale

- Décomposition spectrale de Σ_k:

$$\Sigma_k = Q_k \Delta_k Q_k^t,$$

avec:

- Q_k la matrice des vecteurs propres de Σ_k et $Q_k^{-1} = Q_k^t$
- Δ_k la matrice diagonale des valeurs propres de Σ_k:

$$\Delta_k = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Re-paramétrisation du modèle de mélange

Hypothèse sur Δ_k:

$$\Delta_k = \begin{pmatrix} a_{k1} & 0 & & & & 0 \\ 0 & \ddots & 0 & & & \\ & 0 & a_{kd_k} & & & \\ & & & b_k & 0 & \\ & & & 0 & \ddots & 0 \\ 0 & & & & 0 & b_k \end{pmatrix} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} d_k \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} p - d_k \end{array}$$

Paramètres du modèle

Pour chaque classe k:

- π_k, proportion
- μ_k, vecteur moyenne
- d_k, nombre de dimensions intrinsèques de la classe k
- a_{kj}, $j^{ème}$ valeur propre de Σ_k $(j \in \{1, \dots d_k\})$
- b_k, «bruit» de la classe
- $\tilde{Q}_k$: seulement d_k premiers vecteurs propres

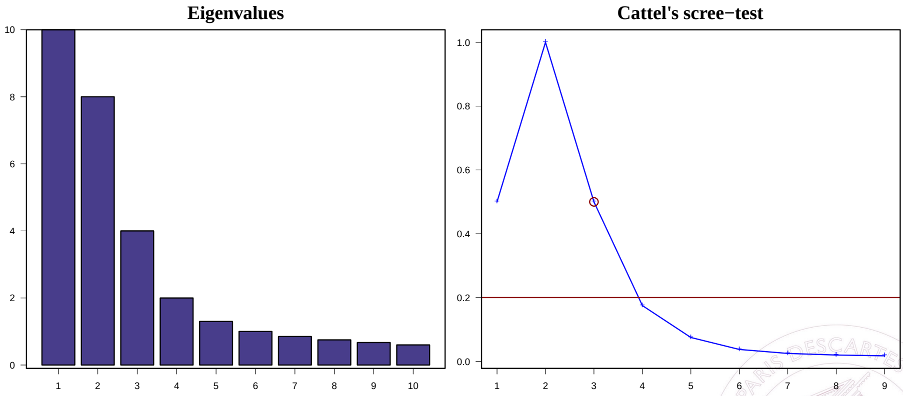
Le modèle $[a_{kj}b_kQ_kd_k]$ et ses sous-modèles

Le modèle général peut être régularisé:

- Au sein de la classe:
  - $a_{k1} = \cdots = a_{kd_k} = a_k$
- Entre les classes:
  - $d_1 = \cdots = d_K = d$
  - $Q_1 = \cdots = Q_K = Q$
  - $b_1 = \cdots = b_K = b$
  - $a_{11} = \cdots = a_{K1} = a_1$

$\Longrightarrow$ 14 modèles

Estimation de d_k



$\Longrightarrow$ on peut faire varier le seuil

- d_k peut être sélectionné par le BIC

Estimation du nombre de groupes K

C'est un problème difficile. Néanmoins différents critères existent:

- BIC (Bayesian Information Criterion)
- ICL (Information Classification Likelihood)
- Heuristique de pente

Plan

Package HDclassif: inputs

La fonction `hhdc` prends en entrée:

- `model`: Le nom d'un des 14 modèles. Par défaut "AkjBkQkDk". Peut être un vecteur.
- `K`: Le nombre de classes. Peut être un vecteur.
- `threshold`: Le seuil pour le scree-test de Cattell. Peut être un vecteur.
- `criterion`: Le critère de sélection (BIC, ICL ou slope).
- `algo`: L'algorithme à utiliser (EM, SEM ou CEM).
- `init`: Le type d'initialisation (random, kmeans, ...).
- `mc.cores`: Nombre de coeurs à utiliser pour le calcul en paralèlle.

Package HDclassif: autres fonctions

- La fonction `predict` calcule la probabilité d'appartenance d'une observation à chacune des classes en fonction de paramètres estimés par `hddc`.
- La fonction `plot` montre la sélection des dimensions intrinsèques.

Package HDclassif: output

La fonction `hddc` donne en sortie:

- `prms`: L'ensemble des paramètres estimés (π_k, μ_k, d_k,a_{kj}, b_k et $\tilde{Q}_k$).
- `Loglik`, `BIC`, `ICL`, `slope`: La valeur de ces critères pour le modèle.
- `all_results`: Tous les modèles qui ont été estimés.
- `class`: La classe associée à chaque observation.
- `posterior`: La matrice $n \times K$ des t_{ik}.

Exemple en direct

Avec tous les risques inhérents !

Conclusion

HDclassif:

- Permet la classification non-supervisée (et supervisée) de façon efficace en grande dimension
- Est particulièrement efficace quand $n \ll p$
- Est flexible (variété de modèles régularisés)
- Permet une sélection de modèle parallèlisable
- Tout commentaire est bienvenu ☺

Merci !

Merci de votre attention !