# Les mélanges finis comme modèle de classification

**Gérard Govaert**

**9 mars 2001**

# Plan

- Introduction
- Modèle de mélange en classification
- Mélange Gaussien
- Exemples
  - Donnés acoustiques
  - Données binaires spatiales
  - Données manquantes
- Conclusion

# Introduction

## Classification automatique

- Données : $n$ individus mesurés par $d$ variables
- Objectif : déterminer une partition ou une hiérarchie formée de classes **homogènes** des $n$ individus
- Classification : 2 significations :
  - Classification automatique, clustering ou classification non supervisée
  - discrimination ou classification supervisée

- Comment ? : optimiser un critère souvent défini à partir d'une métrique
- Exemples : la méthode des k-means et la méthode de Ward optimisent le critère d'inertie intra-classe

$$\sum_{k=1}^{g} \sum_{i \in P_k} d^2(\mathbf{x}_i, \mathbf{g}_k)$$

- Difficultés : choix de la métrique, du critère, choix du nombre de classes, validitation des résultats, ...
- Souvent les données ne sont qu'un échantillon et les résultats obtenus sont étendus à toute la population : pb d'inférence

## Approches probabilistes

- Définition de la notion de « classe naturelles » en utilisant des modèles probabilistes
- Objets à classifier : échantillon issu d'un vecteur aléatoire
- Recherche d'une structure de classes en analysant la densité de ce vecteur aléatoire
- Plusieurs approches
  - Recherche de mode (Wishart, 1969, 1973)
  - Recherche de classes de fortes densités (Hartigan, 1975, Wong, 1982, Wong and Lane, 1983)
  - Utilisation de graphes aléatoires
  - Vraisemblance du lien (Lerman)
  - Mélange fini de probabilités

## Mélange fini de probabilités

- Approche probabiliste qui paraît la plus naturelle :
  chaque classe est modélisée par une distribution de probabilité

$$i \in k \Longrightarrow \mathbf{x}_i \sim \mathcal{L}_k$$

- Références:
  - Scott and Symons, 1971
  - Marriott, 1974
  - Symons, 1981
  - McLachlan and Basford, 1988

# Modèle de mélange en Classification

- Matrice de données $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ n vecteurs de $I\!R^d$
- Matrice de classification $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ n vecteurs de $I\!R^q$ vérifiant $z_{ik} \in \{0, 1\}$ et $\forall i, \sum_{k=1}^{g} z_{ik} = 1$
- Données complètes : $(\mathbf{x}, \mathbf{z})$

| | | | | | | |
|---|---|---|---|---|---|---|
| 3.5 | 2.3 | 0.3 | 4.2 | 0 | 0 | 1 |
| 2.2 | 1.4 | 2.9 | 1.3 | 0 | 1 | 0 |
| 4.2 | 1.7 | 2.2 | 1.1 | 0 | 0 | 1 |
| 2.5 | 2.3 | 0.3 | 4.2 | 0 | 0 | 1 |
| 9.2 | 2.4 | 2.9 | 1.3 | 1 | 0 | 0 |
| 6.2 | 1.2 | 2.2 | 1.1 | 1 | 0 | 0 |

- Pb : trouver $\mathbf{z}$ à partir de $\mathbf{x}$

## Principe

- Connaissant
  - les proportions $p_1, \ldots, p_g$
  - les distributions de chaque classe

  les données sont générées suivant le mécanisme suivant :
  - $\mathbf{z}$ : chaque individu est rangé dans une classe suivant la loi multinomiale de paramètres $p_1, \ldots, p_g$
  - $\mathbf{x}$ : chaque $\mathbf{x_i}$ suit la loi de probabilité associée à la classe à laquelle il appartient

- Plus formellement, cela revient à supposer que les $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont issues d'un vecteur aléatoire de $\mathbb{R}^d$ de loi de probabilité

$$f(\mathbf{x}, \Phi) = \sum_{k=1}^{g} p_k f_k(\mathbf{x}; \theta_k)$$

où

  - $f_k \in$ famille paramétrée de loi de probabilité sur $\mathbb{R}^d$
  - $\Phi = (p_1, \ldots, p_g, \theta)$ paramètre du modèle :
    * $p_1, \ldots, p_g$ proportions du mélange
    * $\theta = (\theta_1, \ldots, \theta_g)$ paramètres des lois de probabilités de chaque composant

## Problème posé

- Retrouver le composant dont est issu chaque $\mathbf{x}_i$
- Le nombre de classe $k$ sera généralement supposé connu
- Deux approches possibles :
  - **Approche vraisemblance**
    - Estimation de $\Phi$ : $\widehat{\Phi}$
    - Estimation de $\mathbf{z}$ : $\mathrm{MAP}(t_{ik}(\widehat{\Phi}))$
  - **Approche vraisemblance classifiante :**
    - Estimation simultanée de $\Phi$ et $\mathbf{z}$

## Vraisemblance

$$L(\Phi) = L(\Phi; \mathbf{x}) = \sum_{i=1}^{n} \log\{\sum_{k=1}^{g} p_k f_k(\mathbf{x}_i; \theta_k)\}$$

## Vraisemblance complétée (ou classifiante)

$$L_C(\mathbf{z}, \Phi) = L(\Phi; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log\{p_k f_k(\mathbf{x}_i, \theta_k)\}$$

## Algorithme CEM

- Estimations de $\mathbf{z}$ et $\Phi$ : maximisation de $L_C$
- Algorithmes généralement proposés : optimisation alternée de $L_C$
- Présentation de ces algorithmes en s'appuyant sur EM :

$$CEM = EM + \text{étape de classification}$$

## Algorithme CEM

- Étape 0 : solution initiale $(\Phi^0, \mathbf{z}^0)$
- Étape E : calcul des $t_{ik} = \frac{p_k^c f_k(\mathbf{x}_i;\theta_k^c)}{\sum_{\ell=1}^{g} p_\ell^c f_k(\mathbf{x}_i;\theta_\ell^c)}$
- Étape C : $\mathbf{z}^{c+1} = \mathrm{MAP}(t_{ki})$
  Les $t_{ki}$ sont remplacées par des 1 ou des 0
- Étape M : maximisation de $L_C(\mathbf{z}^{c+1}, .)$
  - $p_k^{c+1} = \frac{\# z_k^c}{n}$
  - $\theta_k^{c+1}$ : estimation max. de vrais. en utilisant la classe $k$ comme échantillon

Exemple du modèle de mélange gaussien :

$$\mu_k^{c+1} = \frac{1}{\# z_k^{c+1}} \sum_{i \in z_k^{c+1}} \mathbf{x}_i$$

$$\Sigma_k^{c+1} = \frac{1}{\# z_k^{c+1}} \sum_{i \in z_k^{c+1}} (\mathbf{x}_i - \mu_k^{c+1})(\mathbf{x}_i - \mu_k^{c+1})'$$

CEM : algorithme de classification très général

## Lien entre EM et CEM

- Matrice de classification $\mathbf{z} = (z_{ih})$ ou $\mathbf{c} = (c_{ih})$
  - Dure : $z_{ik} \in \{0, 1\}$ et $\forall i, \sum_{k=1}^{g} z_{ik} = 1$
  - Floue : $c_{ik} \in [0, 1]$ et $\forall i, \sum_{k=1}^{g} c_{ik} = 1$
- Expression de la vraisemblance classifiante

$$L_C(\mathbf{z}, \Phi) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log p_k f_k(\mathbf{x}_i; \theta_k) = \ln f(\mathbf{x}, \mathbf{z}; \Phi)$$

- Extension aux matrices de classification floue

$$L_C(\mathbf{c}, \Phi) = \sum_{i=1}^{n} \sum_{k=1}^{g} c_{ik} \log p_k f_k(\mathbf{x}_i; \theta_k)$$

- Maximiser $L_C(\mathbf{c}, \Phi)$ avec $\mathbf{c}$ floue conduit à une classification dure

- Vraisemblance classifiante floue

$$L_{CF}(\mathbf{c},\Phi) = L_C(\mathbf{c},\Phi) + E(\mathbf{c}) \qquad \text{où } E(\mathbf{c}) = -\sum_{i=1}^{n}\sum_{k=1}^{g} c_{ih} \ln c_{ik}$$

- Résultats d'Hathaway
  - Algorithme d'optimisation alternée de $L_{CF}(\mathbf{c},\Phi)$ = EM
  - $L(\Phi) = L_{CF}(\mathbf{t}(\Phi),\Phi) = L_C(\mathbf{t}(\Phi),\Phi) + E(\mathbf{t}(\Phi))$ où $\mathbf{t}(\Phi)$ : matrice de classification floue définie par les $t_{ik}(\Phi)$

## Conclusions

- EM = algorithme de classification floue
- Si $\mathbf{t}(\Phi)$ très séparée : $L(\Phi) \approx L_C(P, \Phi)$
- Différence entre EM et CEM : présence du terme d'entropie
- Approche classification : recherche de composants assez séparés
- Comparaison avec le critère de Bezdek
  - $\sum_{i=1}^{n} \sum_{k=1}^{g} c_{ik} \ln p_k f_k(\mathbf{x}_i; \theta_k) + E(\mathbf{c})$
  - $\sum_{i=1}^{n} \sum_{k=1}^{g} c_{ik}^{r} \ln p_k f_k(\mathbf{x}_i; \theta_k)$

# Mélange gaussien

## Le modèle

$$f(\mathbf{x};\Phi) = f(\mathbf{x};p_1,\ldots,p_k,\theta_1,\ldots,\theta_k) = \sum_{k=1}^{g} p_h f k(\mathbf{x};\theta_k)$$

où

$$f_k(\mathbf{x};\theta_k) = f_k(\mathbf{x};\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)\}$$

avec

- $\mu_k$ vecteur moyenne
- $\Sigma_k$ matrice variance

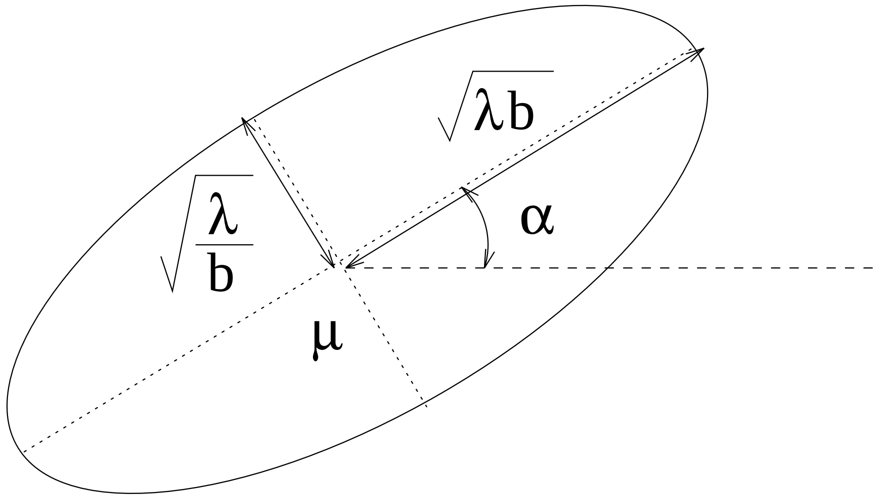
## Paramétrisation de la matrice de variance

Décomposition proposée par Banfield et Raftery (1993) modifiée par Celeux et Govaert (1995)

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

- $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ nombre positif : **volume** de la classe $k$
- $A_k$ matrice diagonale de déterminant 1 avec des valeurs allant en décroissant : **forme** de la classe $k$
- $D_k$ matrice orthogonale des vecteurs propres : **orientation** de la classe $k$

## Exemple dans le plan




$$B = \begin{bmatrix} b & 0 \\ 0 & 1/b \end{bmatrix} \qquad D = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

Finalement les paramètres du modèle de mélange gaussien sont :

- les centres des classes $\mu_1, \ldots, \mu_g$
- les proportions $p_1, \ldots, p_g$
- les volumes $\lambda_1, \ldots, \lambda_g$
- les formes $A_1, \ldots, A_g$
- les orientations $D_1, \ldots, D_g$

## Volume et proportion

Taille d'une classe : deux notions souvent confondues

- Proportion : cardinal de la classe
- Volume : place occupée par la classe

On peut avoir :

- Classe de faible volume et de grande proportion
- Classe de grand volume et de faible proportion

## Variantes du modèle de mélange gaussien

En jouant sur différentes caractéristiques du modèle :

- Formes sphériques ($I$), diagonales ($B$) ou quelconques ($C$)
- Proportions, volumes, formes et directions identiques ou non suivant les classes

plusieurs variantes peuvent être proposées :

- $[\Pi, \lambda I]$ : modèle le plus simple
- $[\Pi, \lambda_k I]$
- $[\Pi, \lambda B]$
- ...
- $[\Pi_k, \lambda_k C_k]$ : modèle le plus compliqué (aucune contrainte)

## Algorithme CEM dans le cas gaussien

**Étapes E et C : construction de $\mathbf{z}$ à partir de $\Phi$**

Rappel : chaque $\mathbf{x}_i$ est rangé dans la la classe qui maximise

$$t_{ik} = \frac{p_k f_k(\mathbf{x}_i; \theta_k)}{\sum_{\ell=1}^{g} p_\ell f_\ell(\mathbf{x}_i; \theta_\ell)}$$

Cela revient à minimiser

$$d^2_{\Sigma_k^{-1}}(\mathbf{x}_i, \mu_k) + \ln|\Sigma_k| - 2\log p_k$$

## Étape M : construction de $\Phi$ à partir de z

Recherche de $\Phi$ maximisant

$$-\frac{1}{2}\sum_{k=1}^{g}\sum_{i\in k}(\mathbf{x_i}-\mu_k)'\Sigma_k^{-1}(\mathbf{x_i}-\mu_k)-\frac{1}{2}\sum_{k=1}^{g}\#k\log|\Sigma_k|+\sum_{k=1}^{g}\#k\log p_k$$

On obtient

- $\mu_k = \overline{\mathbf{x}}_k$ centre de gravité de la classe $k$
- $p_k = \frac{\#z_k}{n}$ si modèle $[P_k]$ et $p_k = \frac{1}{g}$ si modèle $[P]$
- $\Sigma_k$ minimise

$$F(\Sigma_k) = \sum_{k=1}^{g}\mathrm{trace}(W_k\Sigma_k^{-1}) + \sum_{k=1}^{g}\#k\ln|\Sigma_k|$$

où $W_k = \sum_{i\in k}(\mathbf{x_i}-\overline{\mathbf{x}}_k)(\mathbf{x}-\overline{\mathbf{x_i}}_k)'$

## Modèle $[P, \lambda I]$

- Étapes E, C : affectation des individus aux classes avec la distance

$$d^2(\mathbf{x}_i, \mu_k)$$

- Étape M :
  - $\mu_k = \overline{\mathbf{x}}_k$ centre de gravité de la classe $k$
  - $p_k = \frac{1}{g}$
  - $\lambda$ : mimimisation de

$$F(\lambda) = \frac{1}{\lambda}\mathrm{trace}(W) + nd\ln(\lambda) \qquad \text{où } W = \sum_{k=1}^{g} W_k$$

$$\lambda = \frac{\mathrm{trace}(W)}{nd}$$

- Expression de la vraisemblance

$$F(\lambda) = nd + nd \ln \frac{\mathrm{trace}(W)}{nd}$$

- Maximisation de la vraisemblance classifiante = minimisation du critère d'inertie intra-classe trace($W$)
- Algorithme CEM = algorithme des centres mobiles
- Conclusion : prendre le critère d'inertie revient à supposer que les classes sont :
  - sphériques
  - de même proportion
  - de même volume

## Modèle $[P, \lambda_k I]$

- Étapes E, C : Affectation des individus aux classes avec la distance

  $\frac{1}{\lambda_k} d^2(\mathbf{x}_i, \mu_k) + d \ln(\lambda_k)$ (surfaces séparatrices : hypersphères)

- Étape M :
  - $\mu_k = \overline{\mathbf{x}}_k$ centre de gravité de la classe $k$
  - $p_k = \frac{1}{g}$
  - $\lambda_k$ : mimimisation de

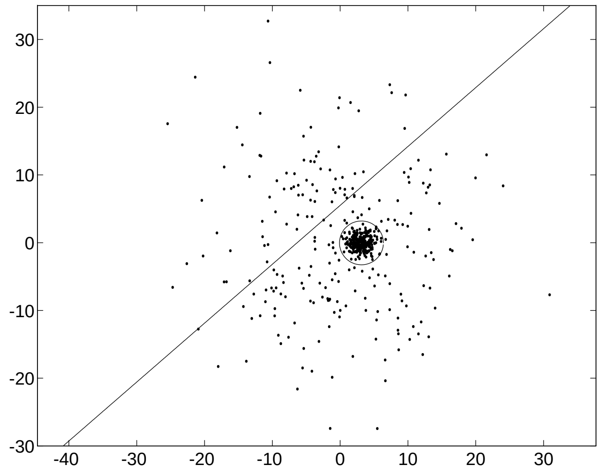
$$F(\lambda_1, \ldots, \lambda_g) = \sum_{k=1}^{g} \frac{1}{\lambda_k} \mathrm{trace}(W_k) + d \sum_{k=1}^{g} \# P_k \ln(\lambda_k)$$

$$\lambda_k = \frac{\mathrm{trace}(W_k)}{\# P_k}$$

- Expression de la vraisemblance

$$F(\lambda_1, \ldots, \lambda_g) = \sum_{k=1}^{g} \ln \frac{\operatorname{trace}(W_k)}{\#P_k}$$

Ce modèle permet de classifier sans aucune difficulté les données




Sans modèle de mélange : difficile de proposer cette distance et ce critère à partir seulement d'une interprétation métrique

## Modèle $[\Pi, \lambda B]$

- Étapes E, C : Affectation des individus aux classes avec la distance

$$d^2_{B^{-1}}(\mathbf{x}_i, \mu_k) \qquad \text{(distance euclidienne pondérée)}$$

- Étape M :
  - $\mu_k = \overline{\mathbf{x}}_k$ centre de gravité de la classe k et $p_k = \frac{1}{g}$
  - $\lambda$ et $B$: mimimisation de

$$F(\lambda, B) = \frac{1}{\lambda}\mathrm{trace}(WB^{-1}) + nd\ln(\lambda)$$

$$B = \frac{\mathrm{diag}(W)}{|\mathrm{diag}(W)|} \qquad \text{et} \qquad \lambda = \frac{|\mathrm{diag}(W)|^{1/p}}{n}$$

- Expression de la vraisemblance

$$F(\lambda, B) = nd + n \ln |\mathrm{diag}(W)| + \text{ cste}$$

- Maximisation de la vraisemblance classifiante = minimisation du critère $|\mathrm{diag}(W)|$
- Méthode permettant de trouver simultanément :
  - partition
  - pondération des variables

## Modèle $[\Pi, \Sigma]$

- Étapes E, C : Affectation des individus aux classes avec la distance

$$d^2_{\Sigma^{-1}}(\mathbf{x}_i, \mu_k)$$

- Étape M :
  - $\mu_k = \overline{\mathbf{x}}_k$ centre de gravité de la classe k
  - $p_k = \frac{1}{g}$
  - $\Sigma$ : mimimisation de

$$F(\Sigma) = \mathrm{trace}(W\Sigma^{-1}) + nd\ln(\Sigma)$$

$$\Sigma = \frac{W}{n}$$

- Expression de la vraisemblance

$$F(\Sigma) = nd + nd \ln |W| - nd \ln n$$

- Critère $|W|$ quelquefois proposé, à juste titre, pour obtenir des classes :
  - non sphériques
  - mais toutes de même forme

## Résumé

| Modèle | distance | critère | Remarques |
|---|---|---|---|
| $\mathbf{p}, \lambda I$ | $d^2(\mathbf{x}_i, \mu_k)$ | $\mathrm{trace}(W)$ | CEM = k-means |
| $\mathbf{p}, \lambda_k I$ | $\frac{1}{\lambda_k} d^2(\mathbf{x}_i, \mu_k) + d \ln(\lambda_k)$ | | surf. separ. : hypersph. |
| $\mathbf{p}, \lambda B$ | $d^2_{B^{-1}}(\mathbf{x}_i, \mu_k)$ | $\mathrm{diag}(W)$ | classification + pond. |
| $\mathbf{p}, \Sigma$ | $d^2_{\Sigma^{-1}}(\mathbf{x}_i, \mu_k)$ | $\|W\|$ | |

# Exemples d'utilisation

## Avantages de l'utilisation des mélanges pour la classification

- Prendre en compte de nombreux aspects particuliers des données dans un cadre unique
- Permet d'analyser les critères de classification classiques ($tr(W)$, $|W|$,...) (hypothèses sous-jacentes)
- Généralise des approches classiques : permet de proposer et de justifier de nouveaux critères de classification
- Existence d'algorithmes associés souples et puissants (EM et sa version classificatoire CEM)
- Type des variables
  - Continue : mélange Gaussien
  - Binaires : mélange de Bernoulli

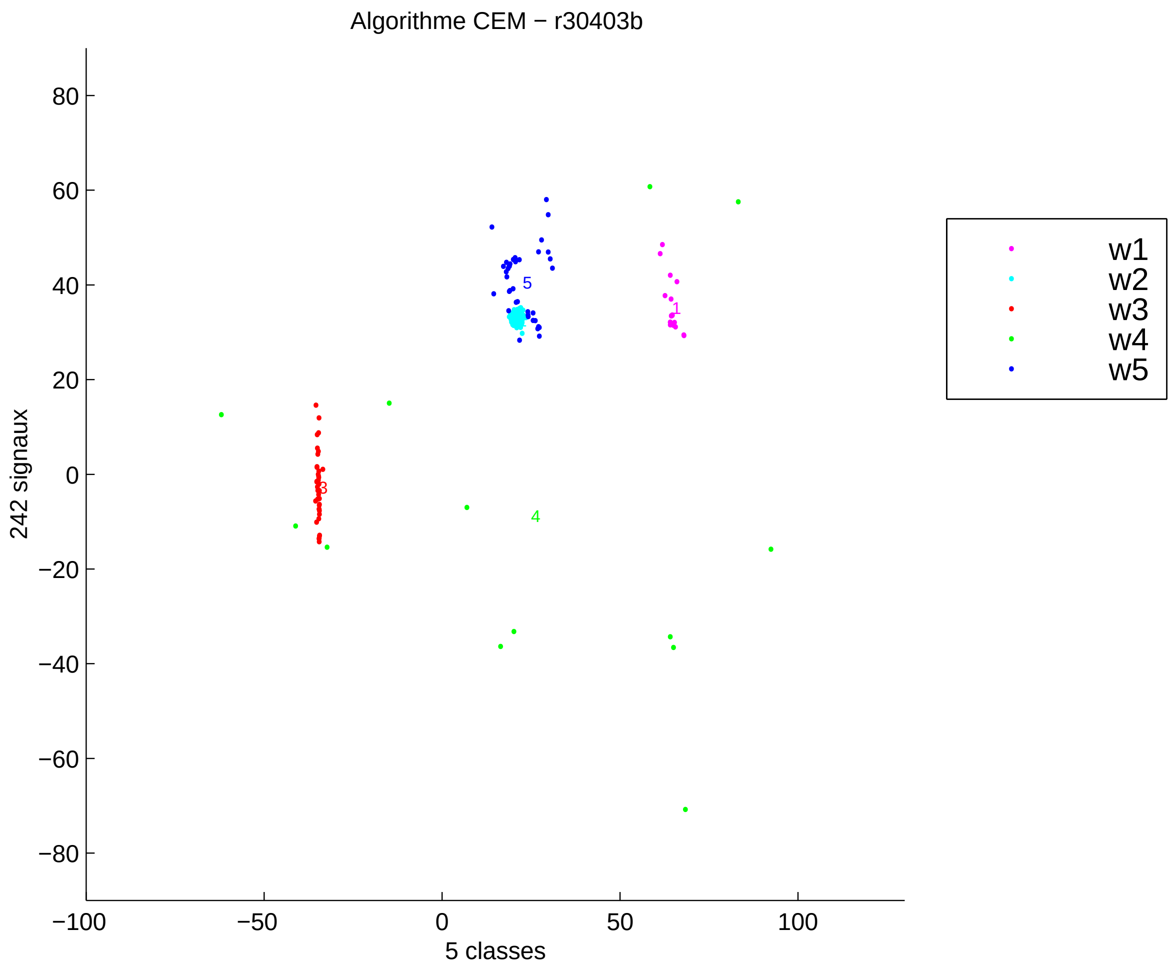
- Choix du nombre de classes : choix de modèles
- Données manquantes
- Echantillon non iid : modèle de markov caché (données spatiales)
- Classes connues et inconnues : discrimination et clustering
- ...

Nous allons illustrer ces différents points au travers de quelques exemples d'application

# Données acoustiques

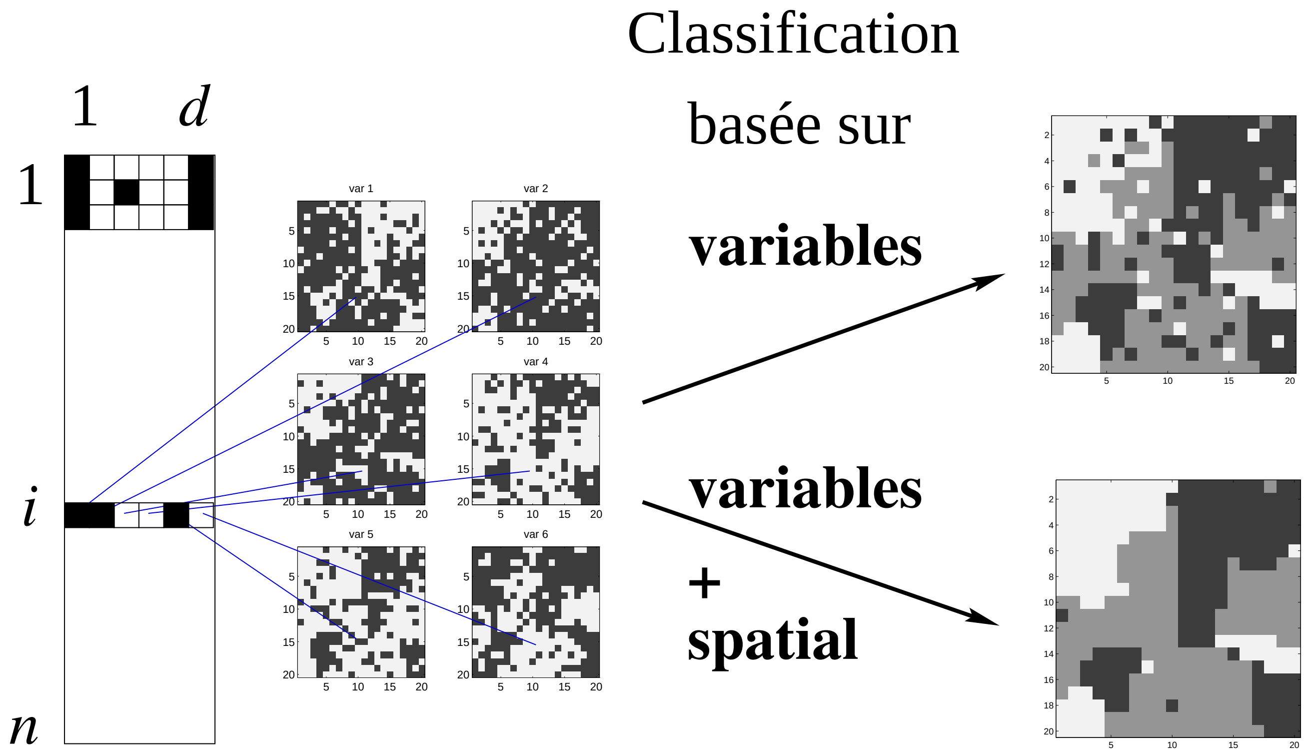
- Probléme : détecter les défauts d'une cuve
- Données : bruits émis à la surface d'une cuve quand celle-ci est soumise à une pression
- Objectifs :
  - détection des zones en défaut
  - arrêter l'expérience avant explosion
- Démarche retenue :
  - Former des classes de défaut
  - Décider s'il s'agit de classes de défaut ou non

- Choix du modèle retenu
  - Classes Gaussiennes : $\lambda_k A_k$
  - Classe de bruit
  - Proportions très différentes : $p_k$
  - Nombre de classes : ICL


Algorithme CEM − r30403b
242 signaux
5 classes
80
60
40
20
0
−20
−40
−60
−80
−100
−50
0
50
100
5
1
3
4
w1
w2
w3
w4
w5

# Données binaires et spatiales

350 parcelles, 22 présence d'animaux




$n$ observations localisées

$d$ variables binaires

partition en

$k$ classes

## Classification binaire

Voir présentation sur les données discrètes

## Classification spatiale

- Information supplémentaire : position géographique de chaque individu
- Plusieurs solutions
  - Hiérarchie avec contiguïté (Legendre 1987)
  - Position spatiale comme variables supplémentaires (Berry 1966)
  - Filtrage
  - Champ de Markov caché

## Champ de Markov caché

- Extension naturelle du modèle de mélange
- Le modèle Les données sont générées suivant le mécanisme suivant :
  - $\mathbf{z}$ : la répartition spatiale des classes $\sim$ champ de Markov
  - $\mathbf{x}$ : chaque $\mathbf{x_i}$ suit la loi de probabilité associée à la classe à laquelle il appartient
- Remarque : toute la partie spatiale du modèle est prise en compte uniquement dans la définition des classes

**Classification cachée z :**

$p(\mathbf{z};\beta) = \exp\left(\beta \sum_{i\sim i'} \mathbf{z}_i^T \cdot mathbf z_{i'}\right)/W(\beta)$

Champ de Markov

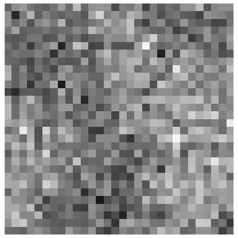
$\beta$ coefficient de régularité spatiale

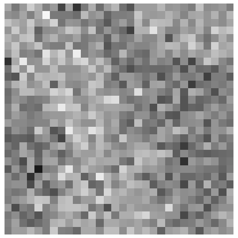
Graphe de voisinage : notion de points voisins

**Observations x :**

$p(\mathbf{x}|\mathbf{z};\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{z}_i;\theta)$

et $\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$





Remarques

- Les sites ne sont pas nécessairement sur une grille
- Il suffit d'avoir le graphe de voisinage
- Possibilité de pondérer les relations de voisinage

## Algorithmes NEM et NCEM

- Estimation des paramètres
  - Maximisation de la vraisemblance : EM difficile à mettre en place
  - Approximations : méthode de Chalmont, NEM
- Clustering
  - ICM : suppose les paramètres connus
  - Vraisemblance classifiante : NCEM

## Algorithme NEM : Neighborhood EM

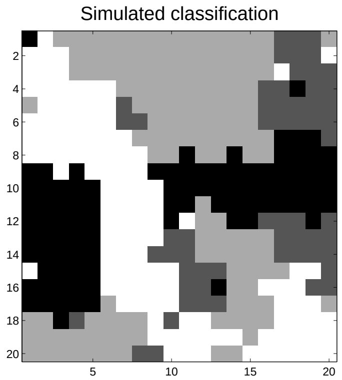
- Critère
$$W(\mathbf{c},\theta) = FCL(\mathbf{c},\theta) + \beta \sum_{i\sim i'} \mathbf{c}_i \cdot \mathbf{c}_{i'}^t$$
Pénalisation du critère d'Hathaway $FCL(\mathbf{c},\theta) = \sum_{i=1}^n \sum_{k=1}^g c_{ik} \log \pi_k p_k(\mathbf{x}_i; \mathbf{a}_k) - \sum_{i=1}^n \sum_{k=1}^g c_{ik} \log c_{ik}$ avec un terme de régularité spatiale

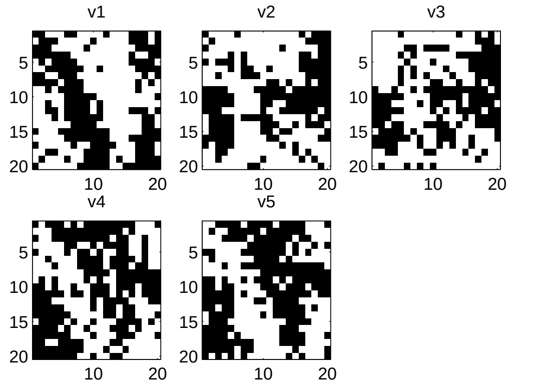
- Optimisation alternée
  - Etape E : $\mathbf{c}^{(n+1)} = \arg\max_{\mathbf{c}} W(\mathbf{c}, \Phi^{(n)})$ : méthode itérative de point fixe
  - Etape M : identique à l'étape M de EM

- Approximation de champ moyen

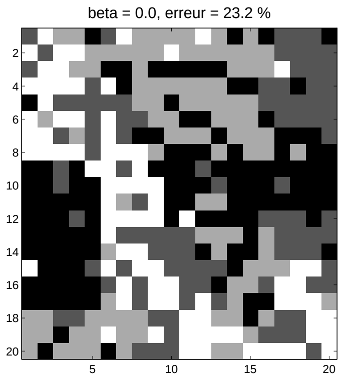
# Exemple de données binaires spatiales

Partition simulée | Observations simulées

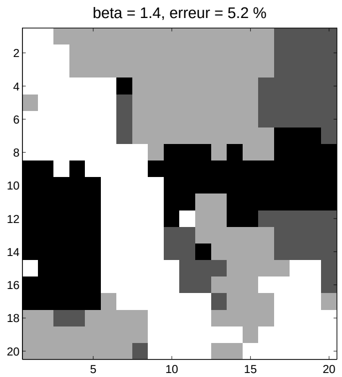


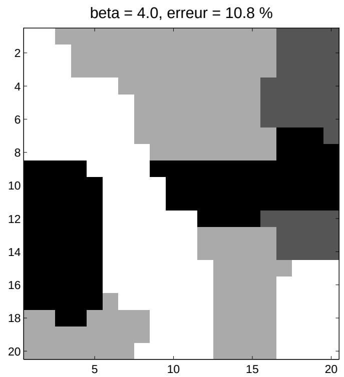

Classifications par NEM :




$\beta = 0, P_E = 23\%$




$\beta = 1.4, P_E = 5\%$




$\beta = 4, P_E = 11\%$

Discussion

- Modèle probabiliste adapté aux données binaires
- Régularisation spatiale sans contrainte de connexité
- Classification floue obtenue $\simeq$ probabilités appartenance a posteriori (distribution a priori de champ de Markov)
- Critère optimisé calculable $\rightarrow$ choix entre optimums locaux
- Estimation $\beta$

# Données manquantes

- Individus : stations météorologiques
- Variables : Mesures climatiques définies sur une année :
  - Niveau d'ensoleillement
  - Pluie
  - Température maximale
  - Température minimale
  - ...
- Position géographique de chaque station
- Présence de nombreuses valeurs manquantes
- Objectif : classifier les stations

# Données manquantes

$$
\begin{array}{cc}
 & d \text{ variables} \\
\begin{array}{c} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{array} &
\boxed{\begin{array}{cccc} 2 & 10 & -2 & ? \\ -3 & 11 & ? & -1 \\ & \vdots & & \\ ? & ? & -5 & -9 \end{array}}
\end{array}
\longrightarrow
\begin{array}{cc}
 & k \text{ classes} \\
\begin{array}{c} \mathbf{c}_1^T \\ \mathbf{c}_2^T \\ \vdots \\ \mathbf{c}_n^T \end{array} &
\boxed{\begin{array}{cc} 0.23 & 0.77 \\ 0.93 & 0.07 \\ \vdots & \\ 0.04 & 0.96 \end{array}}
\end{array}
\longrightarrow
\begin{array}{cc}
 & k \text{ classes} \\
\begin{array}{c} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{array} &
\boxed{\begin{array}{cc} 0 & 1 \\ 1 & 0 \\ \vdots & \\ 0 & 1 \end{array}}
\end{array}
$$

| Données | Partition floue | Partition dure |
|---|---|---|
| $\mathbf{x}^o = \mathbf{x}_1^{o_1}, \ldots, \mathbf{x}_n^{o_n}$ | $\mathbf{c} = \mathbf{c}_1, \ldots, \mathbf{c}_n$ | $\mathbf{z} = \mathbf{z}_1, \ldots, \mathbf{z}_n$ |

## Solutions

- Supprimer les observations incomplètes
- Remplacer les valeurs manquantes par la moyenne correspondante
- Utiliser le modèle statistique

## Hypothèses

- Si $\mathbf{r} = (r_{ij})$ avec $r_{ij} = \begin{cases} 1 \text{ si } x_{ij} \text{ observé} \\ 0 \text{ sinon} \end{cases}$
- Absence aléatoire (MAR) :

$$p(\mathbf{r}|\mathbf{x}, \mathbf{z}) = p(\mathbf{r}|\mathbf{x}^o)$$

## Vraisemblance

- Maximiser la log-vraisemblance associée $L(\theta) = \ln\left(\prod_{i=1}^{n} p(\mathbf{x_i^0};\theta)\right)$
- Algorithme EM

$$\theta^{(cn+1)} = \arg\max_{\theta}\Big(Q(\theta,\theta^{(n)}) = E\big[\log p(\mathbf{x}^o,\mathbf{X}^m,\mathbf{Z};\theta)\big|\mathbf{x}^o;\theta^{(n)}\big]\Big)$$

  - Etape E : calcul des $t_{ik}^{(n+1)} = p(Z_{ik}=1|\mathbf{x}^o;\theta^{(n)})$

$$t_{ik}^{(n+1)} = \frac{\pi_k p_k^{o_i}(\mathbf{x}_i^{o_i};\mu_k^{(n)},\Sigma_k^{(n)})}{\sum_{\ell=1}^{q}\pi_\ell p_\ell^{o_i}(\mathbf{x}_i^{o_i};\mu_\ell^{(n)},\Sigma_\ell^{(n)})}$$

  - Etape M

$$\mu_{kj}^{(n+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(n+1)}\left(r_{ij}x_{ij} + (1-r_{ij})\mu_{kj}^{(n)}\right)}{\sum_{i=1}^{n} c_{ik}^{(n+1)}} = \alpha\,\mu_{kj}^{(n+1)} + (1-\alpha)\,\mu_{kj}^{(n)}$$

## Vraisemblance Classifiante

- Maximiser la vraisemblance classifiante associée

$$CL(P,\theta) = \ln p(\mathbf{x}^{\mathbf{0}}, \mathbf{z}; \theta)$$

- Algorithme CEM
  - Etape E : identique à l'algorithme précédent
  - Etape C : MAP
  - Etape M :

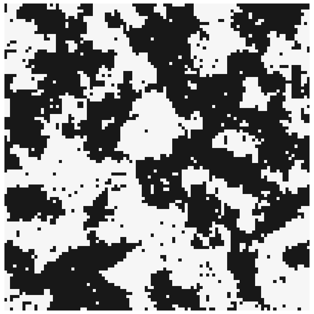
$$\mu_{kj}^{(n+1)} = \frac{\sum_{i=1}^{n} c_{ik}^{(n+1)} r_{ij} x_{ij}}{\sum_{i=1}^{n} c_{ik}^{(n+1)} r_{ij}}$$

## Données manquantes dans le cas spatial

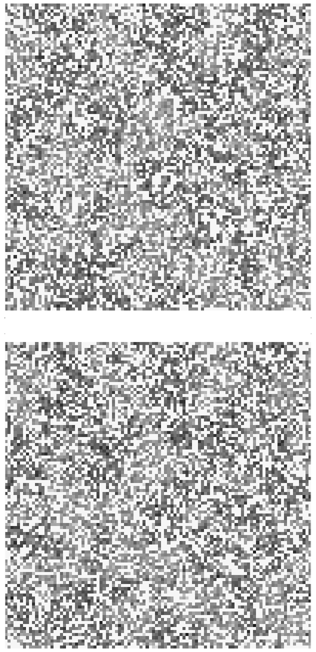
- Idée : reprendre simultanément NEM ou NCEM avec données manquantes
- Exemple de résultats sur données simulées
  - taille $n = 20 \times 20$ et $100 \times 100$
  - $\mathbf{z}^{sim} \sim$ champ de Markov $k = 2$ classes, $\beta = 1$
  - $\mathbf{x}|\mathbf{z}^{sim}$, $\mathbf{x}_i \in \mathbb{R}^2, \mathbb{R}^6$ et $\mathbb{R}^{10}$
    $\mu_1 = (0, \ldots, 0)$; $\mu_2 = (1, \ldots, 1)$; $\Sigma_1 = \Sigma_2 = \lambda.I$
    3 niveaux de bruit : $\lambda = 0.1\ d, 0.2\ d$, et $d$
  - $\mathbf{r}$ : absence complètement aléatoire
    proportions d'absence : $\{0\%, 10\%, \ldots, 90\%\}$

$$\xrightarrow{\text{Gibbs}} \mathbf{z}^{sim} \xrightarrow{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})} \mathbf{x} \xrightarrow{\text{Absence}} \mathbf{x}^o \xrightarrow{\text{NEM}} \hat{\mathbf{z}}$$

## Exemple



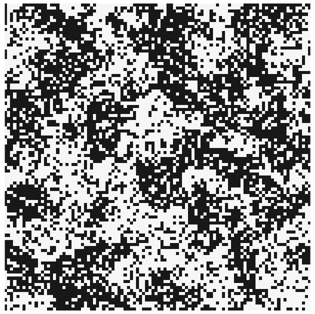
$\mathbf{z}^{sim}$



$\mathbf{x}^{o}$

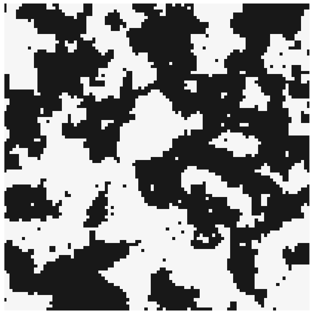
$\lambda = 0.2\ d$

40 % manquant



$\hat{\mathbf{z}}_{\beta=0}$

erreur = 22 %



$\hat{\mathbf{z}}_{\beta=1}$

erreur = 7 %

# Conclusion

- EM ou CEM ?
  - Biais dans l'estimation des paramètres
  - Inconsistance de l'estimation (nombre de paramètres à estimer augmentant avec $n$)
  - Introduit un aspect classificatoire dans le modèle de mélange : recherche de classes assez séparées
  - Passerelle entre la classification automatique et les modèles statistiques
- Nombre de classes
- Choix de modèles et du nombre de classes

# References

[1] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[2] B. J. L. Berry. Essay on commodity flows and the spatial structure of the indian economy. Research paper 111, Departement of Geography, University of Chicago, 1966.

[3] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[4] J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[5] P. Legendre. Constrained clustering. *Develop. Numerical Ecology. NATO ASI*, 14:289–307, 1987.

[6] F. H. C. Marriott. Separating mixtures of normal distributions. *Biometrics*, 31:767–769, 1975.

[7] G.J. McLachlan and K.E. Basford. *Mixture Models, Inference and applications to clustering*. Marcel Dekker, New York, 1988.

[8] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[9] M. J. Symons. Clustering criteria and multivariate normal mixture. *Biometrics*, 37:35–43, march 1981.

[10] D. Wishart. Mode analysis. In A. J. Cole, editor, *Numerical Taxonomy*, pages 282–308, New York, 1969. Academic Press.

[11] M. A. Wong. A hybrid clustering method for identifying high-density clusters. *J. Am. Statis. Assoc.*, 77:841–847, 1982.

[12] M. A. Wong and T. lane. A k-th nearest neighbour lustering procedure. *Journal of the Royal Statistical Society (series B)*, 45(3):362–368, 1983.