Modèle de mélange et classification

Gérard Govaert

Heudiasyc
CNRS et université de technologie de Compiègne
gerard.govaert@utc.fr

8 Décembre 2008

Plan

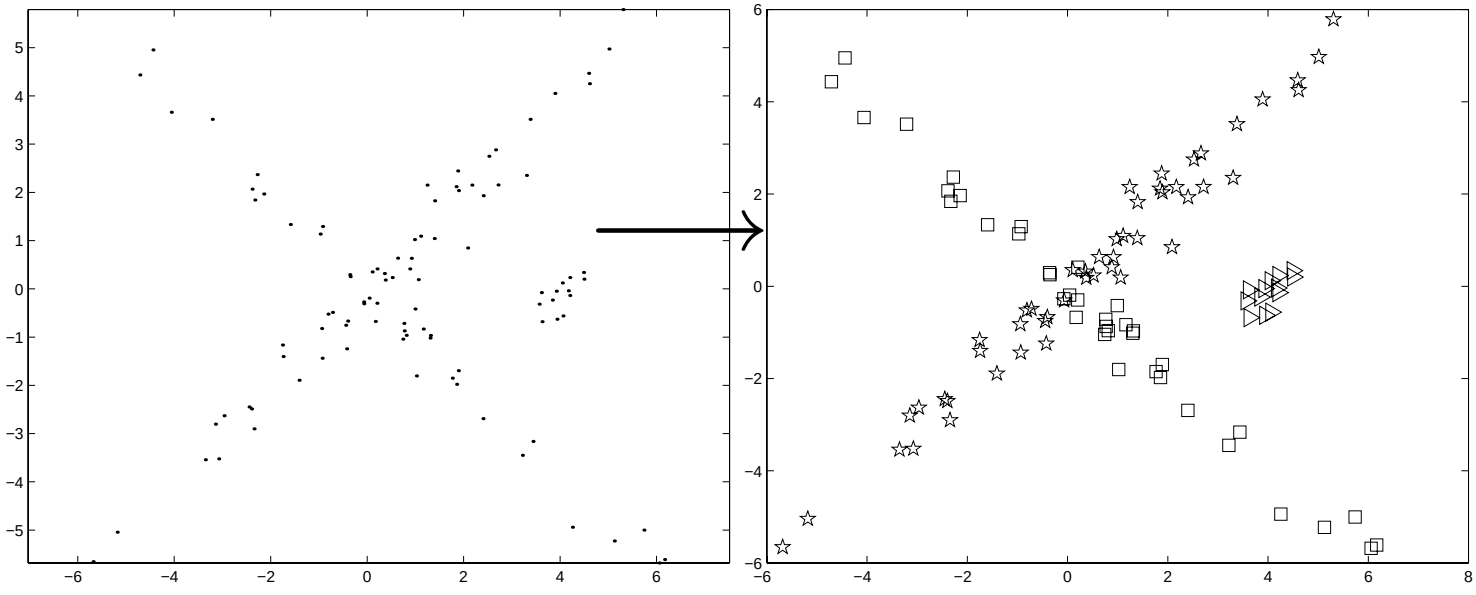
- Les approches probabilistes de la classification
- Le modèle de mélange
- Utilisation du modèle de mélange en classification
- Mélange gaussien multivarié
- Mélange multinomial multivarié

Première partie I

Les approches probabilistes de la classification

Objectif de la classification non supervisée

- Trouver une partition dans un jeu de données . . .

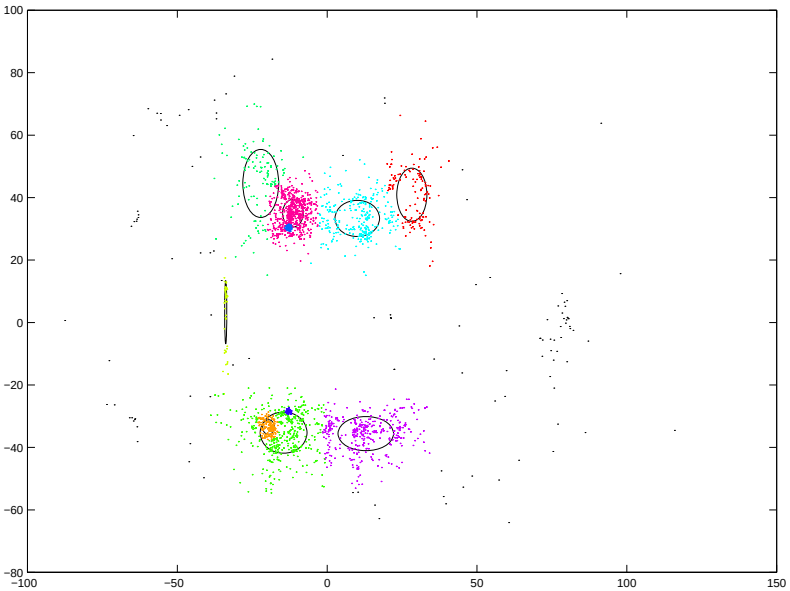


. . . afin de synthétiser des données complexes et volumineuses

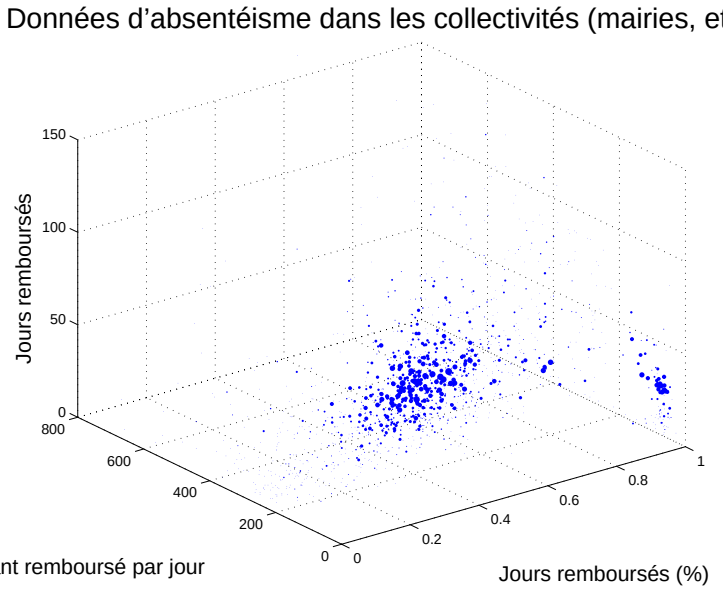
- Objectifs de la classification automatique : organiser un ensemble d'objets ou d'individus en classes homogènes

Exemples de classification non supervisée

Industrie : contrôle de cuve



Sociologie : absences d'employés



Une bonne partition est composée de classes interprétables :

- Une classe de craquements indique un défaut dans la cuve
- Une classe d'employés suggère une cause d'absence similaire

Approches classiques de la classification non supervisée

- Choix d'un critère géométrique :
  - Critère d'inertie intraclasse
- Recherche d'une structure de classique optimisant ce critère :
  - Algorithme des centres-mobiles (k-means)
  - Classification hiérarchique ascendant de Ward

Critère d'inertie intra-classe

$$C(P) = \sum_k \sum_{i \in P_k} \|x_i - \bar{x}_k\|^2_{\mathbf{M}}$$

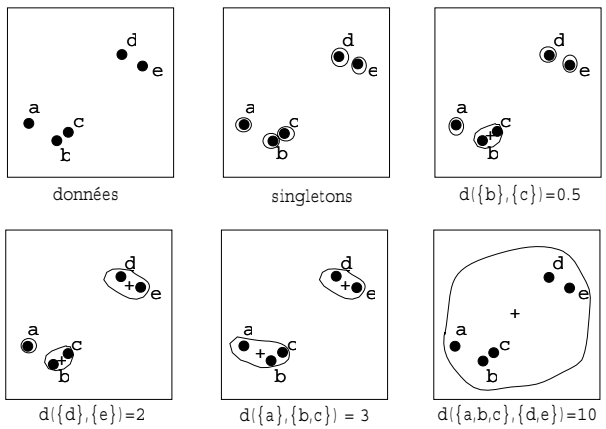
où

- $\|\cdot\|_{\mathbf{M}}$ est la distance euclidienne définie par la matrice $\mathbf{M}$ dans $\mathbb{R}^d$
- $\bar{x}_k$ est la moyenne de la classe P_k :

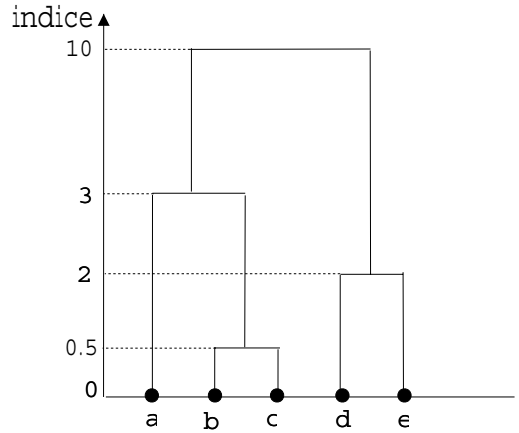
$$\bar{x}_k = \frac{1}{n_k} \sum_k \sum_{i \in P_k} x_i$$

avec n_k cardinal de la classe P_k

Classification hiérarchique de Ward



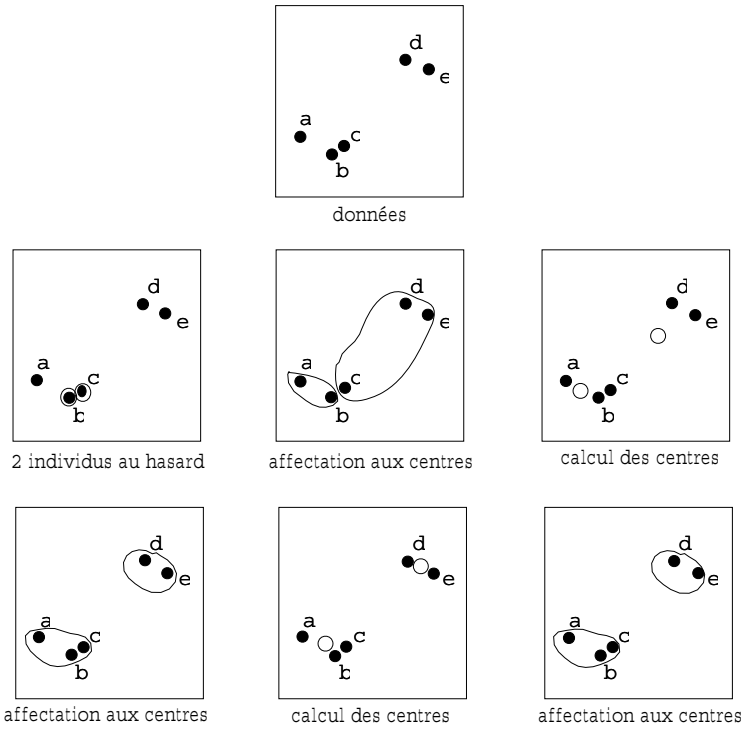
Classification hiérarchique ascendante
(méthode de Ward)



Dendogramme

- Optimisation sous-optimale du critère d'inertie intra-classe
- Une partition est obtenue en coupant le dendrogramme

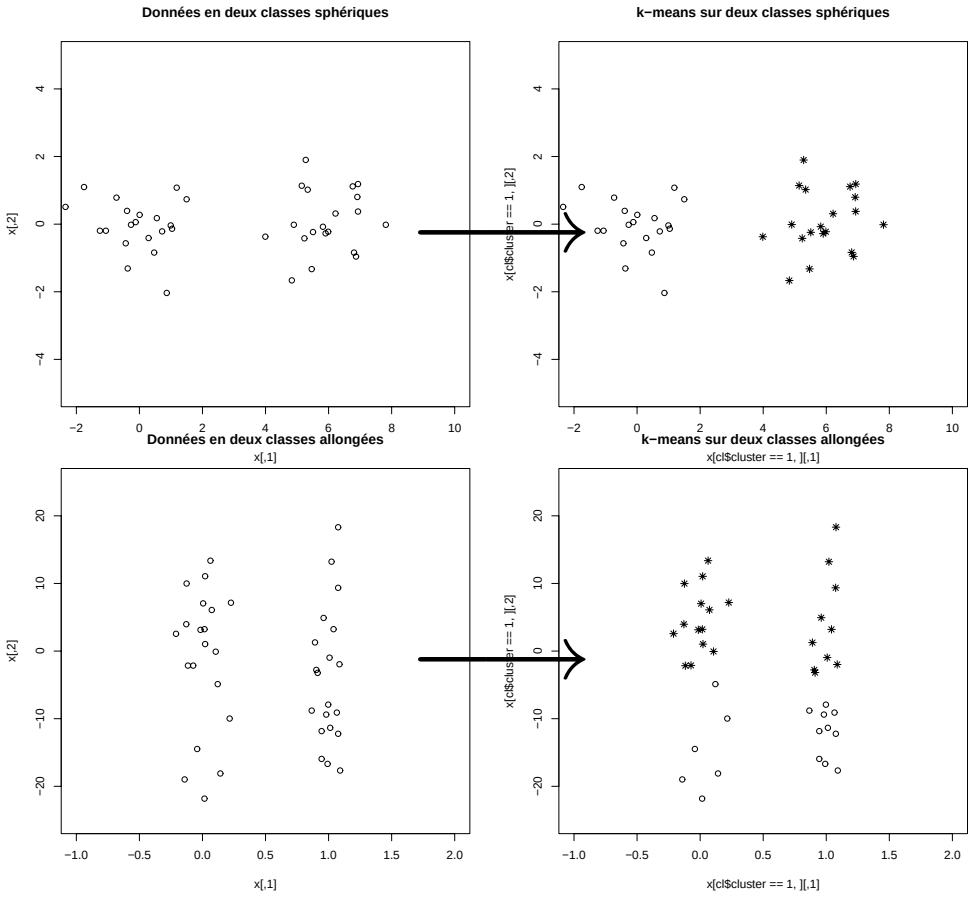
Algorithme des centres mobiles



Algorithme des centres mobiles

Optimisation alternée entre la partition et le centre des classes

La métrique **M** identité : un choix courant mais risqué



Difficultés

- Choix de la métrique
- Choix du critère
- Choix du nombre de classes : le critère diminue quand le nombre de classes augmente
- Sélection d'un algorithme
- Extension des résultats à une population de référence

Question de fond : problème mal posé . . .

Qu'est-ce qu'une classe ?

Solution : passer d'une démarche géométrique à une démarche probabiliste

- Utilisation des modèles probabilistes de classification pour formaliser l'idée intuitive de la notion de classe naturelle
- Evolution de l'approche algorithmique, heuristique et géométrique vers une approche plus statistique
- Avantages :
  - Analyse précise et interprétation statistique de certains critères métriques dont les différentes variantes n'étaient pas toujours bien claires : trace(S_W) et $|S_W|$
  - Définir de nouvelles variantes répondant à des hypothèses précises
  - Cadre formel pour proposer des solutions à des problèmes difficiles : nombre de classes, validation des résultats, ...

Cadre général

- Objets à classifier : échantillon d'un vecteur aléatoire
- Classification obtenue en analysant la densité de ce vecteur
- Différentes approches probabilistes de la classification
  - Approches paramétriques
  - Approches non paramétriques

Approches non-paramétriques

- Aucune hypothèse sur la distribution de probabilités
- S'appuyer sur la forme de cette distribution :
  - Classes de forte densité : exemple de Hartigan (1975)
    - Sous-ensemble connexe de points de densité $>$ à un certain seuil
    - En faisant varier ce seuil, il obtient un arbre hiérarchique de classes
  - Multimodalité :
    - Recherche des maxima de la densité
    - Classes modales obtenues en affectant les individus à ces maxima
- Préalable : estimation de la densité
  - Histogramme
  - Méthode des plus proches voisins
  - Méthode des noyaux

Approches paramétriques

- Hypothèses sur la distribution de probabilité induisant une classification et formalisant ainsi la notion de classes « naturelles »
- Modèle de mélange :
  - Objet de ce chapitre
  - Modèle paramétrique le plus utilisé en classification automatique
- Autres modèles :
  - Modèles fonctionnels à effet fixe
  - Processus ponctuel en statistique spatiale

Modèles fonctionnels à effet fixe

- Forme : données = structure + erreur
  - Structure inconnue mais fixe
  - Erreur aléatoire
- Application à la classification :
  - Exemple simple : $\boldsymbol{x}_i = \boldsymbol{y}_i + \varepsilon_i$
    - $\boldsymbol{y}_i$ appartient à un ensemble de K centres $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$
    - $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$
  - Exemple sur des données de similarité : $d(a, b) = \delta(a, b) + \varepsilon(a, b)$
    - δ est une distance ultramétrique
    - La CAH du lien moyen : maximum local de la vraisemblance de ce modèle lorsque l'erreur est gaussienne (Degens, 1983)

Processus ponctuel en statistique spatiale

- Exemples : la répartition des arbres dans une forêt ou des étoiles dans l'espace
- Certains de ces processus correspondent à une organisation en agrégats : modèles probabilistes associés à une classification
- Exemple : processus de Neyman-Scott
  1. K points $\mathbf{a}_1, \ldots, \mathbf{a}_K$ sont tirés au hasard suivant une distribution uniforme sur une région convexe
  2. Les tailles $n_1, \ldots, n_K$ des classes sont tirées au hasard, par exemple à l'aide d'une distribution de Poisson
  3. Pour chaque classe k, n_k points sont tirés au hasard en utilisant une distribution sphérique centrée en $\mathbf{a}_k$ (ex : loi gaussienne de moyenne $\mathbf{a}_k$)

Deuxième partie II

Modèles de mélange

Le modèle

La variable Z

Utilisations du modèle de mélange

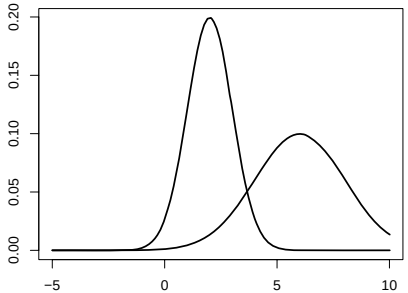
Les hypothèses du modèle de mélange

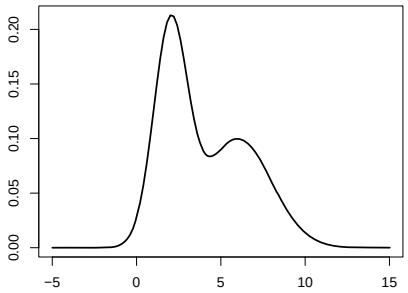
- Le modèle de mélange fini de lois de probabilité consiste à supposer que les données proviennent d'une source contenant plusieurs sous-population.
- Chaque sous-population est modélisée de manière séparée.
- La population totale est un mélange de ces sous-populations.
- Le modèle résultant est un modèle de mélange fini.

Définition d'un modèle de mélange

La forme générale d'un modèle de mélange à g composant est

$$f(x) = \sum_k \pi_k f_k(x)$$





- π_k : proportions du mélange
- $f_k(.)$: densités des composants

La paramétrisation des densités des composants dépend de la nature (continue ou discrète) des données observées.

Modèle à structure cachée

- Le modèle de mélange est un modèle à données incomplètes
- Les données complétées sont

$$\mathbf{y} = (\mathbf{x}, \mathbf{z}) = (y_1, \ldots, y_n) = ((x_1, z_1), \ldots, (x_n, z_n))$$

où les données manquantes sont $\mathbf{z} = (z_1, \ldots, z_n) = (z_{ik})$

  - z_i = composant de i
  - $z_{ik} = 1$ si i provient du groupe k et 0 sinon

3.5	2.3	0.3	4.2	1
2.2	1.4	2.9	1.3	2
4.2	1.7	2.2	1.1	3
2.5	2.3	0.3	4.2	3

3.5	2.3	0.3	4.2	1	0	0
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

$\mathbf{z}$ définit une partition $P = (P_1, \ldots, P_g)$ des données observées $\mathbf{x}$ avec $P_k = \{i \mid z_{ik} = 1\}$

Modèle génératif

Connaissant

- les proportions $\pi_1, \ldots, \pi_g$ et
- les distributions f_k des composants,

les données sont générées suivant le schéma suivant

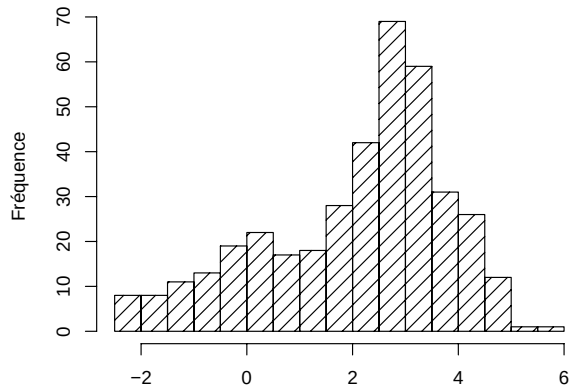
- $z_i \sim \mathcal{M}(1, \pi_1, \ldots, \pi_g)$ (distribution multinomiale)
- $\boldsymbol{x}_i \sim$ distribution de densité f_{z_i}

Réalité de la variable Z ?

- Passereaux
  - L'existence de deux groupes, qui ont une signification physique, est incontestable
  - Seule inconnue : valeur de cette variable pour les individus de l'échantillon
- Utilisation du modèle de mélange dans des situations où l'existence même d'une telle variable n'est pas sûre
- Exemple d'une étude portant sur les programmes de vaccination contre les oreillons

Exemple de la vaccination

- Log-concentration d'anticorps de 385 enfants non vaccinés contre les oreillons
- Mode important autour de 3
- Second mode (moins net) autour de 0
- Pour ces données, il est connu qu'une population homogène aurait dû conduire à une distribution sensiblement gaussienne
- Explication raisonnable des deux modes : mélange de deux groupes
  - enfants immunisés naturellement
  - enfants non immunisés
- Groupes moins séparés
- Existence de deux groupes : hypothèse de travail suggérée par les données et non directement observable



Nombre de composants K

- Peut être connu (par exemple pour les passereaux où la notion de composant a une signification physique bien précise)
- Plus généralement K est inconnu et doit être estimé : paramètre supplémentaire
- Problème difficile
- Plusieurs critères de sélection ont été proposés

Utilisations du modèle de mélange

- Modélisation de populations hétérogènes
- Développement d'estimateurs robustes : modèle de contamination, élément atypique,...
- Estimation de densité semi-paramétrique
- Classification automatique

Troisième partie III

Utilisation du modèle de mélange en classification

L'approche ML

L'approche CML

Comparaison de deux approches

Lien avec la classification floue

Introduction

- Les modèles de mélange sont de plus en plus utilisés en classification automatique
- Pourquoi ?
  - Idée intuitive d'une population composée de plusieurs classes
  - Liens forts avec des méthodes de références comme l'algorithme des *k-means*
  - Capacité de traiter de manière assez naturelle de nombreuses situations particulières
- Comment ?
  - Approche ML (*maximum likelihood*)
  - Approche CML (*classification maximum likelihood*)

Principe de l'approche ML

- Estimation des paramètres du mélange par la méthode du maximum de vraisemblance :
- Détermination de la partition en rangeant chaque individu dans la classe la plus probable conditionnellement à cet estimation : MAP
  - Calcul des t_{ik}, probabilités conditionnelles que les observation $\boldsymbol{x}_i$ proviennent de la classe k en utilisant les paramètres estimés
  - Affectation de chaque observation à la classe qui maximise t_{ik}

Estimation du maximum de vraisemblance

- Problème posé : maximisation de la vraisemblance

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \prod_i \sum_k \pi_k f_k(\boldsymbol{x}_i; \alpha_k)$$

ou, de manière équivalente de la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_i \log \left(\sum_k \pi_k f(\mathbf{x}_i, \alpha_k) \right)$$

  - $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \alpha_1, \ldots, \alpha_K)$: paramètre du modèle
  - $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$: échantillon
  - $f_k(., \alpha_k)$: densité du k^{e}composant
- Équations de vraisemblance ne possédant généralement pas de solution analytique
- Utilisation d'algorithmes itératifs

Un algorithme itératif

- On peut montrer que, si le paramètre α_k est un vecteur de nombres réels α_k^r, la solution de ces équations de vraisemblance doit vérifier :

$$\pi_k = \frac{1}{n}\sum_i t_{ik} \quad \forall k \quad \text{et} \quad \sum_i t_{ik} \frac{\partial \log f_k(\mathbf{x}_i, \alpha_k)}{\partial \alpha_k^r} = 0 \quad \forall k, r \quad (1)$$

$$\text{avec} \quad t_{ik} = \frac{\pi_k f_k(\mathbf{x}_i, \alpha_k)}{\sum_\ell \pi_\ell f_\ell(\mathbf{x}_i, \alpha_\ell)} \quad (2)$$

- Ces équations suggèrent l'algorithme itératif suivant :
  1. Initialisation de θ
  2. Calcul des t_{ik} à partir de ce paramètre en utilisant (2)
  3. Remise à jour de θ à partir de ces t_{ik} en utilisant (1) et retour en (2)
- Si cet algorithme converge, alors le point fixe obtenu vérifiera les équations de vraisemblance
- En fait, cet procédure n'est rien d'autre que l'application au modèle de mélange de l'algorithme EM

EM : Données complétées et vraisemblance classifiante

- Algorithme s'appuyant sur la notion de données complétées
- Modèle de mélange :

$$(\mathbf{x}, \mathbf{z}) = (\underbrace{(x_1, \ldots, x_n)}_{\text{données}}, \underbrace{(z_1, \ldots, z_n)}_{\text{labels inconnus}})$$

- Vraisemblance des données complétées ou vraisemblance classifiante :

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \prod_{i,k} (\pi_k f_k(\mathbf{x}_i; \alpha_k))^{z_{ik}}$$

Log-vraisemblance classifiante :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i,k} z_{ik} \log (\pi_k f_k(\mathbf{x}_i; \alpha_k))$$

Principe de l'algorithme EM

- EM s'appuie sur la log-vraisemblance complétée

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{k,i} z_{ik} \log \pi_k f_k(\boldsymbol{x}_i; \alpha_k)$$

plus simple à manipuler que la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i} \log \sum_{k} \pi_k f_k(\boldsymbol{x}_i; \alpha_k)$$

- Maximisation itérative de

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= E(L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})|\mathbf{x}, \boldsymbol{\theta}') = \sum_{i,k} E(z_{ik}|\mathbf{x}, \boldsymbol{\theta}') \log \pi_k f_k(\boldsymbol{x}_i; \alpha_k) \\ &= \sum_{i,k} t_{ik} \log \pi_k f_k(\boldsymbol{x}_i; \alpha_k) \quad \text{« vraisemblance pondérée »} \end{aligned}$$

où $t_{ik} = E(z_{ik}|\mathbf{x}, \boldsymbol{\theta}') = P(z_{ik} = 1|\mathbf{x}, \boldsymbol{\theta}')$ $\rightarrow$ probabilité conditionnelle d'appartenance de $\boldsymbol{x}_i$ au composant k

L'algorithme EM

- **Étape initiale** : θ^0
- Répéter jusqu'à la convergence :
  - **Étape E** : Calcul des probabilités conditionnelles t_{ik} que l'observation i provienne du composant k pour la valeur courante du paramètre du mélange :

$$t_{ik}^m = \frac{\pi_k^m f(\boldsymbol{x}_i; \alpha_k^m)}{\sum_\ell \pi_\ell^m f(\boldsymbol{x}_i; \alpha_\ell^m)}$$

  - **Étape M** : Mettre à jour l'estimation des paramètres en maximisant l'espérance de la vraisemblance des données complétées. Cela conduit à pondérer, pour le composant k, l'observation i avec la probabilité conditionnelle t_{ik}.
    - $\pi_k^{m+1} = \frac{1}{n} \sum_i t_{ik}^m$
    - α_k^{m+1} : résolution d'équations de vraisemblance

Caractéristiques de EM

- EM fait croître la vraisemblance à chaque itération
- Sous certaines conditions, il converge vers l'unique solution consistante des équations de vraisemblance
- Facile à programmer
- Peu gourmant en place mémoire
- Bon comportement pratique
- Situation de convergence lente (en particulier, lorsque les composants sont très mélangés)
- Des maxima nombreux et même des points selles
- Très populaire : voir le livre de McLachlan et Krishnan (1997)

Exemple d'un mélange gaussien de $\mathbb{R}$ à 2 composants

- Initialisation de π_1, π_2, μ_1, μ_1, σ_1^2 et σ_2^2
- Étape E : calcul des t_{ik} pour $i = 1, \ldots, n$ et $k = 1, 2$

$$t_{ik} = \frac{\pi_k \varphi(x_i; \mu_k, \sigma_k^2)}{\sum_{\ell=1}^{2} \pi_\ell \varphi(x_i; \mu_\ell, \sigma_\ell^2)}$$

- Étape M : si on note $n_1 = \sum_i t_{i1}$ et $n_2 = \sum_i t_{i2}$, on obtient

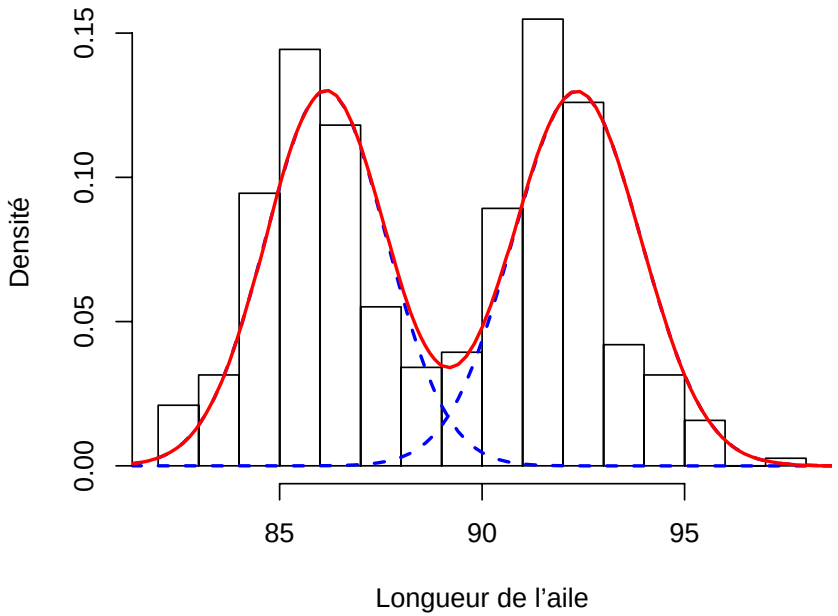
$$\pi_1 = \frac{n_1}{n} \quad \text{et} \quad \pi_2 = \frac{n_2}{n}$$

$$\mu_1 = \frac{1}{n_1} \sum_i t_{i1} x_i \quad \text{et} \quad \mu_2 = \frac{1}{n_2} \sum_i t_{i2} x_i$$

$$\sigma_1^2 = \frac{1}{n_1} \sum_i t_{i1} (x_i - \mu_1)^2 \quad \text{et} \quad \sigma_2^2 = \frac{1}{n_2} \sum_i t_{i2} (x_i - \mu_2)^2$$

Applications aux données passereaux et oreillons

	Paramètres	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2	nb d'itér.
Passereaux	initiaux	0.50	0.50	85	95	1	1	
	obtenus	0.49	0.51	86.1	92.3	2.5	2.2	54
Oreillons	initiaux	0.50	0.50	−3	5	1	1	
	obtenus	0.30	0.70	−0.07	2.98	1.35	0.79	221



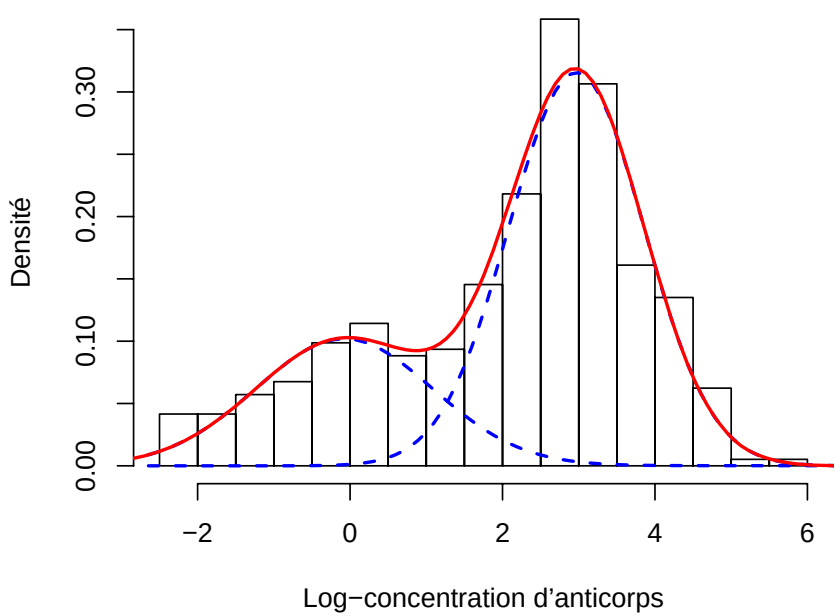
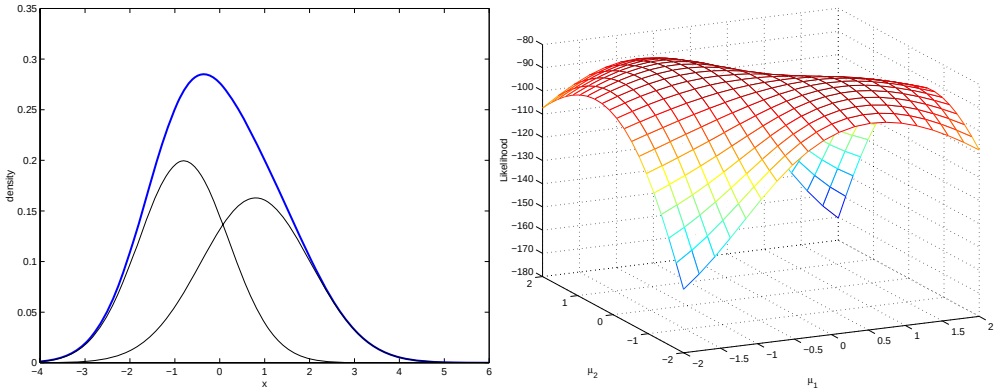
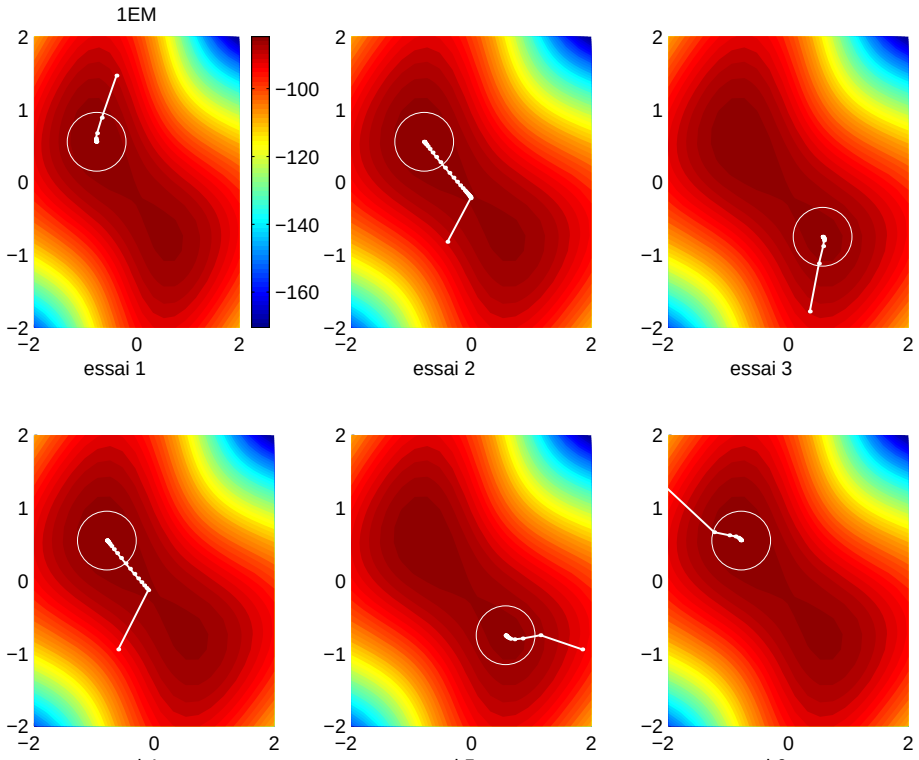


Illustration du problème des maximas locaux





Méthodes d'accélération de EM

- Introduction de méthodes d'optimisation (gradient, Newton-Raphson)
- Inconvénient : perte de la simplicité de l'algorithme EM
- Autres approches
  - Utilisées pour les données de grandes tailles
  - Incremental EM
  - Lazy EM
  - Sparse EM

L'algorithme SEM

- Version stochastique de l'algorithme EM (stochastic EM)
- Ajout d'une étape S de classification aléatoire

- **Étape initiale** : θ^0
- Répéter jusqu'à la convergence :
  - **Étape E** : calcul des t_{ik} comme dans l'algorithme EM
  - **Étape S** : tirage au hasard d'une classe d'affectation pour chaque point suivant la distribution $(t_{ik}, k = 1, \ldots, g)$
  - **Étape M** : Mise à jour des paramètres en maximisant la vraisemblance des données complétées : les estimations du maximum de vraisemblance des π_k et des α_k sont obtenues en utilisant les classes de la partition $\mathbf{z}^{(c+1)}$ comme sous-échantillons
    - Proportions : $\pi_k^{(c+1)} = \frac{n_k^{(c+1)}}{n}$
    - $\alpha_k^{(c+1)}$: dépendent du modèle de mélange retenu

Caractéristiques de SEM

- Pas de convergence au sens habituel
- SEM génère une chaîne de Markov dont la distribution stationnaire est (plus ou moins) concentrée autour de l'estimateur du maximum de vraisemblance
- Estimation à partir de SEM
  - SEMmean : effectuer la moyenne des valeurs obtenues après une période de chauffe
  - SEMmax : retenir la valeur du paramètre ayant conduit à la plus grande vraisemblance

Principe de l'approche CML

- Approche ML : la partition est un sous-produit issu de l'estimation de $\boldsymbol{\theta}$
- Approche CML : la partition $\mathbf{z}$ est ajoutée au paramètre à estimer
- Estimation simultanée des paramètres du mélange et des labels en maximisant la vraisemblance associée : vraisemblance classifiante

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i,k} z_{ik} \log\left(\pi_k f_k(\boldsymbol{x}_i; \alpha_k)\right)$$

- Cette approche revient à rechercher une partition de l'échantillon de telle sorte que chaque classe k soit assimilable à un sous-échantillon issu de la loi $f_k(., \alpha_k)$

Variante : vraisemblance classificante restreinte

$$L_{CR}(\boldsymbol{\theta}, \mathbf{z}) = \sum_{i,k} z_{ik} \log \left(f_k(\boldsymbol{x}_i; \alpha_k) \right)$$

Lien entre les deux vraisemblances classifiantes :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = L_{CR}(\boldsymbol{\theta}, \mathbf{z}) + \sum_k n_k \log \pi_k$$

- n_k est le cardinal de la classe k
- $\sum_k n_k \log \pi_k$: terme de pénalité

Algorithme CEM : Classification EM

- Version classifiante de EM

- **Étape 0** : $\theta^{(0)}$
- Répéter jusqu'à la convergence :
  - **Étape E** : Calcul des probabilités conditionnelles t_{ik} comme dans EM
  - **Étape C** : Affectation de chaque observation i au composant qui maximise la probabilité conditionnelle t_{ik} (MAP)
  - **Étape M** : Mise à jour des paramètres en maximisant la vraisemblance des données complétées : les estimations du maximum de vraisemblance des π_k et des α_k sont obtenues en utilisant les classes de la partition $\mathbf{z}^{(c+1)}$ comme sous-échantillons
    - Proportions : $\pi_k^{(c+1)} = \frac{n_k^{(c+1)}}{n}$
    - $\alpha_k^{(c+1)}$: dépendent du modèle de mélange retenu

Caractéristiques de CEM

- Algorithme itératif faisanr croître à chaque itération la vraisemblance complétée sous des conditions très générales
- Algorithme stationnaire : converge en un nombre fini d'itérations
- De nombreux algorithmes de classification peuvent être présentés comme des cas particuliers de l'algorithme CEM et donc de pouvoir les englober dans une approche probabiliste de la classification :
  - Algorithme des *k-means*

Lien avec les critères métriques

- Le lien avec les *k-means* peut être généralisé
- Proposition : Si le critère de classification se met sous la forme :

$$W(\mathbf{z}, \boldsymbol{\lambda}, D) = \sum_{i,k} z_{ik} D(\mathbf{x}_i, \boldsymbol{\lambda}_k)$$

et s'il existe un réel r t.q. $\int r^{-D(\mathbf{x},\boldsymbol{\lambda})} d\mathbf{x}$ soit indépendante de $\boldsymbol{\lambda}$, alors ce critère est équivalent au critère de vraisemblance classifiante associé à un modèle de mélange de la forme $f(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{s} r^{-D(\mathbf{x},\boldsymbol{\lambda})}$ où s est une constante > 0

Comparaison de deux approches

- CEM, déterminant à chaque itération les paramètres à l'aide d'échantillons tronqués du modèle de mélange, fournit une estimation biaisée
- Estimation inconsistante : le nombre de paramètres à estimer croît avec la taille de l'échantillon
- Il est généralement préférable d'utiliser EM
- Toutefois, lorsque les classes sont bien séparées et les effectifs relativement petits, l'approche classification peut fournir de meilleurs résultats.
- CEM est beaucoup plus rapide que EM : solution lorsqu'il y a des contraintes de temps
  - Temps réel
  - Données de très grande taille

Classification floue

- L'appartenance, vraie ou fausse, d'un objet à une classe est remplacée par un degré d'appartenance
- La matrice de classification $\mathbf{z}$ vérifiant

$$z_{ik} \in \{0, 1\} \quad \text{avec} \quad \sum_k z_{ik} = 1$$

est donc remplacée par une matrice $\mathbf{c}$ vérifiant

$$c_{ik} \in [0, 1] \quad \text{avec} \quad \sum_k z_{ik} = 1$$

- Algorithme le plus connu : *Fuzzy k-means*

Fuzzy k-means : le critère

- *k-means* : déterminer la partition $\mathbf{z}$ minimisant

$$W(\mathbf{z}) = \sum_{i,k} z_{ik} d^2(\mathbf{x}_i, \mathbf{g}_k)$$

  - $\mathbf{g}_k \in \mathbb{R}^p$: point de l'espace (« centre »de la classe)
  - d : distance euclidienne
- Généralisation : déterminer la partition floue $\mathbf{c}$ minimisant

$$W(\mathbf{z}) = \sum_{i,k} c_{ik} d^2(\mathbf{x}_i, \mathbf{g}_k)$$

- Problème : W minimal pour $c_{ik} = 0$ ou $1 \Longrightarrow$ *k-means*
- Solution proposée par Bezdek : minimiser le critère

$$W(\mathbf{c}) = \sum_{i,k} {c_{ik}}^{\gamma} d^2(\mathbf{x}_i, \mathbf{g}_k)$$

où $\gamma > 1$ est 1 coefficient permettant de régler le degré de flou

Fuzzy k-means : l'algorithme

- Répétition des 2 étapes
  1. Centres :
  $$\mathbf{g}_k = \frac{\sum_i c_{ik}^{\gamma} \boldsymbol{x}_i}{\sum_i c_{ik}^{\gamma}}$$
  2. Partition floue :
  $$c_{ik} = \frac{D_i}{\|\boldsymbol{x}_i - \mathbf{g}_k\|^{\frac{2}{\gamma-1}}}$$

  où
  $$D_i = \sum_{\ell} \frac{1}{\|\boldsymbol{x}_i - \mathbf{g}_\ell\|^{\frac{2}{\gamma-1}}}$$
- Valeurs conseillées de γ : intervalle $]1, 2[$

Algorithme EM : un algorithme de classification floue

- L'estimation des paramètres d'un modèle de mélange est une autre façon d'aborder, et de manière plus naturelle, ce problème
- Les t_{ik} définissent une classification floue
- Il est même possible d'aller plus loin et de montrer que l'algorithme EM optimise un critère de classification floue

Interprétation d'Hathaway

- Critère de classification floue : $F_c(\boldsymbol{\theta}, \mathbf{c}) = L_c(\boldsymbol{\theta}, \mathbf{c}) + H(\mathbf{c}) \quad (1)$
  - $L_c(\boldsymbol{\theta}, \mathbf{c}) = \sum_{i,k} c_{ik} \log\left(\pi_k \varphi_k(\boldsymbol{x}_i; \boldsymbol{\alpha})\right)$ vrais. classifiante « floue »
  - $H(\mathbf{c}) = -\sum_{i,k} c_{ik} \log c_{ik}$ fonction d'entropie
- On peut montrer la relation $F_c(\boldsymbol{\theta}, \mathbf{c}) = L(\boldsymbol{\theta}) - KL(\mathbf{c}, \mathbf{t}(\boldsymbol{\theta})) \quad (2)$
- Algorithme d'optimisation alternée du critère F_c :
  1. Maximisation pour $\boldsymbol{\theta}$ fixé : $(2) \Rightarrow \mathbf{c}$ doit minimiser $KL(\mathbf{c}, \mathbf{t}(\boldsymbol{\theta}))$; or cette fonction atteint son minimum 0 pour $\mathbf{c} = \mathbf{t}(\boldsymbol{\theta})$
  2. Maximisation pour $\mathbf{c}$ fixé : $(1) \Rightarrow \boldsymbol{\theta}$ doit maximiser la vraisemblance complétée floue $L_c(\boldsymbol{\theta}, \mathbf{c})$
- On retrouve exactement les deux étapes de l'algorithme *EM*
- Après chaque première étape, on a $F_c(\boldsymbol{\theta}, \mathbf{c}) = F_c(\boldsymbol{\theta}, \mathbf{t}(\boldsymbol{\theta})) = L(\boldsymbol{\theta})$: l'algorithme fait croître la vraisemblance
- EM : algorithme de classification floue

Interprétation de Radford et Neal

- Généralisation à toute situation relevant de l'algorithme *EM*
- Soit $\hat{P}$ une distribution sur l'espace des données complétées et

$$W(\hat{P}, \theta) = E(L_c(\theta, \mathbf{y})|\hat{P}) + H(\hat{P}) \quad (1)$$

- Ceci généralise bien la situation précédente car ici une distribution $\hat{P}$ est définie par un vecteur (c_{ik}) : $E(L_c(\theta, \mathbf{y})|\hat{P}) = L_c(\theta, \mathbf{c})$
- Si on note $P_\theta = P(\mathbf{y}|\mathbf{x}, \theta)$, on peut montrer

$$W(\hat{P}, \theta) = L(\theta) - KL(\hat{P}, P_\theta) \quad (2)$$

- Optimisation alternée du critère W :
  1. Maximisation pour θ fixé : (2) $\Rightarrow$ $\hat{P}$ minimise $KL(\hat{P}, P_\theta) \Rightarrow \hat{P} = P_\theta$
  2. Maximisation pour $\hat{P}$ fixé : (1) $\Rightarrow$ θ maximise $E(L_c(\theta, \mathbf{y})|\mathbf{x}, \hat{P})$
- Ce sont les 2 étapes de *EM*. Par ailleurs, après chaque 1$^{\text{e}}$étape, on a

$$W(\hat{P}, \theta) = W(P_\theta, \theta) = L(\theta)$$

Ce qui montre bien que l'algorithme fait croître la vraisemblance.

Lien entre EM et CEM

- Si on supprime le terme d'entropie du critère F_C, on obtient à chaque étape des partitions « dures »
- Algorithme obtenu : CEM
- Différence entre EM et CEM : présence du terme d'entropie
- Si, à la convergence de EM, les composants sont très séparés, la partition floue $\mathbf{z}(\theta)$ est proche d'une partition et on a
  - $H(\mathbf{z}(\theta)) \approx 0$
  - $L(\theta) = F_C(\theta, \mathbf{z}(\theta)) = L_C(\theta, \mathbf{z}(\theta)) + H(\mathbf{z}(\theta)) \approx L_C(\theta, \mathbf{z}(\theta))$

Quatrième partie IV

Mélange gaussien multivarié

Définition

- Les observations multidimensionnelles $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ dans $\mathbb{R}^d$ sont supposées être un échantillon d'une distribution de probabilité de densité

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_k \pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$$

- Loi conditionnelle φ : loi normale multidimensionnelle

$$\varphi(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

- Paramètre $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K)$
  - Proportions π_k
  - Vecteurs moyennes $\boldsymbol{\mu}_k$
  - Matrices de variance Σ_k

Remarques

- Si on note $d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ la distance quadratique définie par Σ_k^{-1}, φ s'écrit aussi

$$\varphi(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp -\frac{1}{2} d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$$

- Les classes associées aux composants du mélange sont ellipsoïdales, centrées à la moyenne $\boldsymbol{\mu}_k$ et les matrices de variance Σ_k déterminent leurs caractéristiques géométriques
- Dans certaines situations, ce modèle pourra être simplifié en imposant aux proportions d'êtres toutes égales à $\frac{1}{K}$
- Modèle le plus utilisé pour la classification de données quantitatives

Étape E

- Les étapes E des algorithmes EM, SEM et CEM sont identiques
- Pas de problème particulier
- Calcul des des probabilités d'appartenance des $\mathbf{x}_i$ aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} \varphi(\mathbf{x}_i; \boldsymbol{\mu}_k^{(c)}, \Sigma_k^{(c)})}{\sum_\ell \pi_\ell^{(c)} \varphi(\mathbf{x}_i; \boldsymbol{\mu}_\ell^{(c)}, \Sigma_\ell^{(c)})}$$

Étape de classification C

- Etape supplémentaire de classification pour CEM
- La partition $z^{(c+1)}$ est obtenue en rangeant chaque x_i dans la classe maximisant $t_{ik}^{(c)}$:

$$z_{ik}^{(c+1)} = \begin{cases} 1 & \text{si } k = \operatorname{argmax}_k \ t_{ik}^{(c)} \\ 0 & \text{sinon} \end{cases}$$

- Chaque x_i est donc rangé dans la classe qui
  - Maximise $\pi_k \varphi(x_i; \mu_k, \Sigma_k)$
  - Minimise $-\log(\pi_k \varphi(x_i; \mu_k, \Sigma_k))$ ou encore

$$d^2_{\Sigma_k^{-1}}(x_i, \mu_k) + \log|\Sigma_k| - 2\log \pi_k$$

Notations

- Pour unifier la présentation, on utilise la matrice de classification $\mathbf{c} = (c_{ik})$ avec
  - $c_{ik} = t_{ik}^{(c)}$ pour EM (partition floue)
  - $c_{ik} = z_{ik}^{(c)}$ pour CEM et SEM (partition)
- $n_k = \sum_i c_{ik}$
- $S_k = \frac{1}{n_k} \sum_i c_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)'$
- $S_W = \frac{1}{n} \sum_k n_k S_k$
- Lorsque $\mathbf{c}$ correspond à une partition
  - n_k est le cardinal de la classe k
  - S_k est la matrice de variance de la classe k
  - S_W est la matrice de variance intraclasse

Problème posé

- M est donc dans tous les cas la maximisation en θ de

$$L_C(\theta, \mathbf{c}) = \sum_{i,k} c_{ik} \ln \left[\pi_k \varphi(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k)\right]$$

qui s'écrit, à une cste près, dans le cas gaussien multidimensionnel

$$-\frac{1}{2} \sum_{i,k} c_{ik} (\boldsymbol{x}_i - \mu_k)' \Sigma_k^{-1} (\boldsymbol{x}_i - \mu_k) - \frac{1}{2} \sum_k n_k \log |\Sigma_k| + \sum_k n_k \log \pi_k$$

Calcul des proportions et des moyennes

$$-\frac{1}{2}\sum_{i,k} c_{ik}(\boldsymbol{x}_i - \mu_k)'\Sigma_k^{-1}(\boldsymbol{x}_i - \mu_k) - \frac{1}{2}\sum_k n_k \log|\Sigma_k| + \sum_k n_k \log \pi_k$$

- Proportions : maximisation de $\sum_k n_k \log \pi_k$:

$$\pi_k = n_k/n$$

- Moyennes : maximisation de $-\frac{1}{2}\sum_{i,k} c_{ik}(\boldsymbol{x}_i - \mu_k)'\Sigma_k^{-1}(\boldsymbol{x}_i - \mu_k)$:

$$\boldsymbol{\mu}_k = \bar{\boldsymbol{x}}_k = \frac{1}{n_k}\sum_i c_{ik}\boldsymbol{x}_i$$

Calcul des matrices de variance

$$-\frac{1}{2}\sum_{i,k} c_{ik}(\boldsymbol{x}_i - \mu_k)'\Sigma_k^{-1}(\boldsymbol{x}_i - \mu_k) - \frac{1}{2}\sum_k n_k \log|\Sigma_k| + \sum_k n_k \log \pi_k$$

- Les Σ_k doivent alors minimiser la fonction

$$F(\Sigma_1, \ldots, \Sigma_g) = \sum_k n_k \left(\text{trace}(S_k\Sigma_k^{-1}) + \log|\Sigma_k|\right)$$

On obtient $\Sigma_k = S_k$

Valeur atteinte

- Sachant que

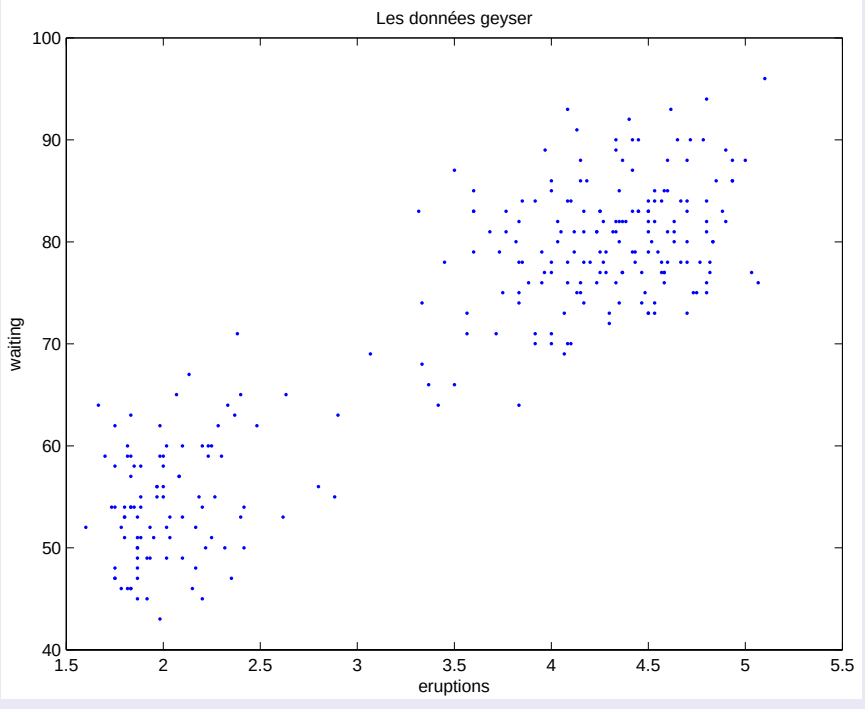
$$L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{1}{2} F(\Sigma_1, \ldots, \Sigma_g) + \sum_k n_k \log \pi_k - \frac{np}{2} \log 2\pi$$

la valeur de L_c maximisée par $\boldsymbol{\theta}$ vérifie

$$L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{1}{2} \sum_k n_k \log |S_k| + \sum_k n_k \log \pi_k - \frac{np}{2} (\log 2\pi + 1)$$

Un exemple : the Old Faithful Geyser

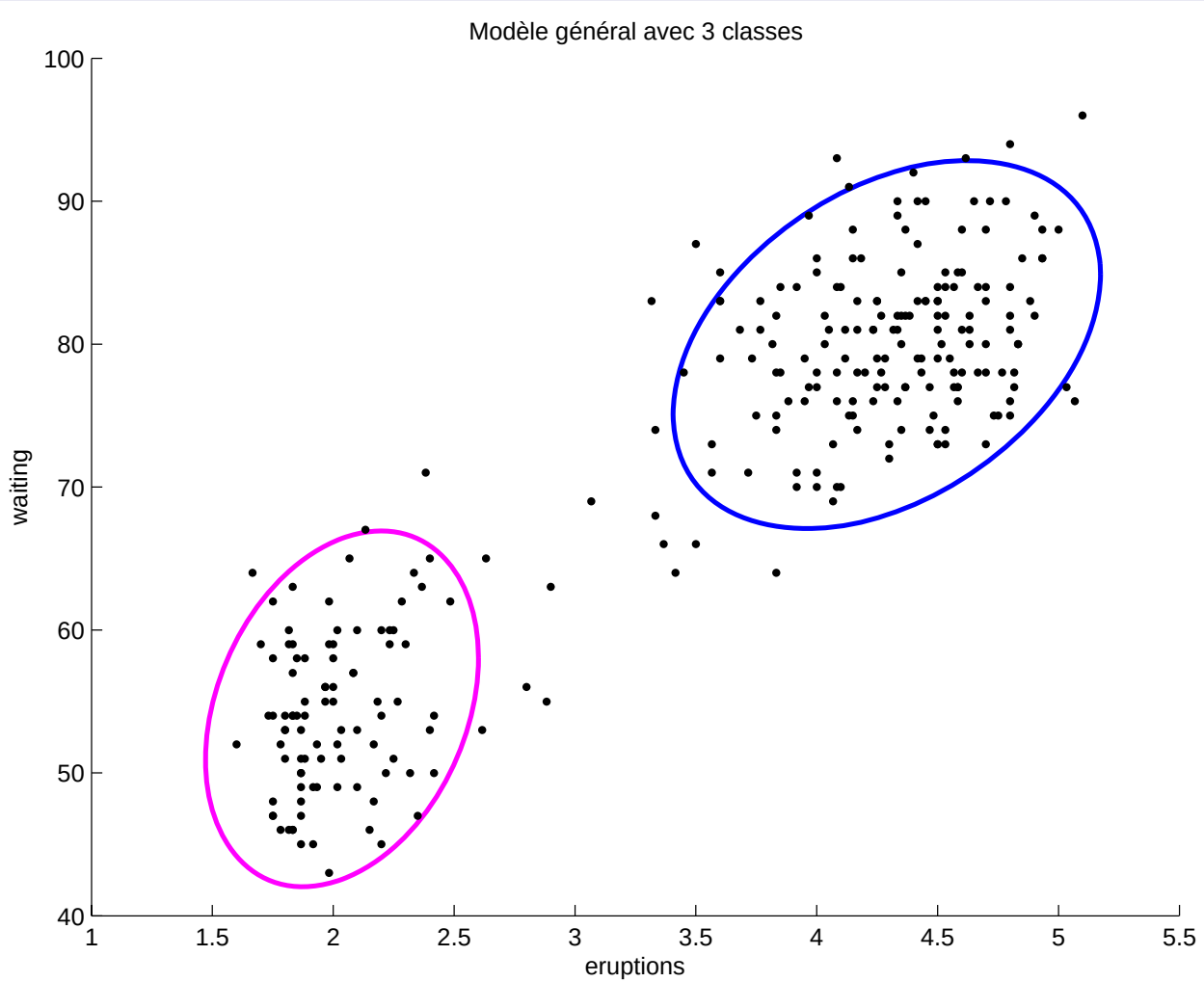
- Eruptions d'un geyser situé dans le parc de Yellowstone aux USA et dénommé Old Faithful car ses éruptions sont très régulières
- Deux variables (durée de l'éruption et temps d'attente avant la suivante) mesurées sur 272 observations



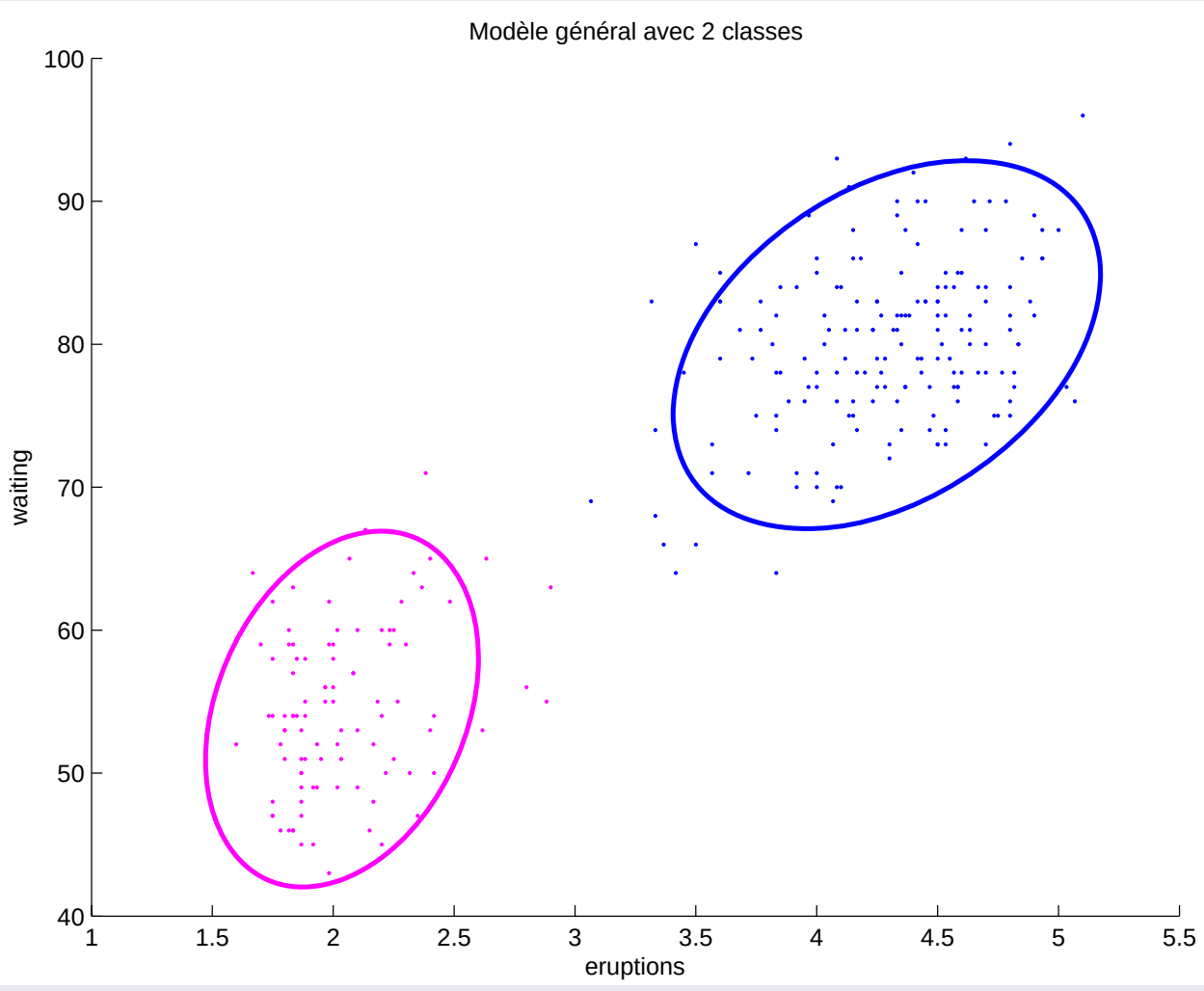
Utilisation du logiciel Mixmod avec 2 classes

```
----------------------------------------------------------------
*               Cluster size : list by cluster
*                                .cluster 1   97
*                                .cluster 2   175
*                Proportions : list by cluster
                                 .cluster 1    0.35588
                                 .cluster 2    0.64412
*                      Means : list by mean vector of clusters
                                 .cluster 1
                                              2.03641      54.4788
                                 .cluster 2
                                              4.28968      79.9684
*                  Variances : list by clusters of variance matrix
                                 .cluster 1
                                              0.0692         0.4354
                                              0.4354        33.6987
                                 .cluster 2
                                              0.1699         0.9402
                                              0.9402         36.0422
*              Log-likelihood : -1130.264
----------------------------------------------------------------
```

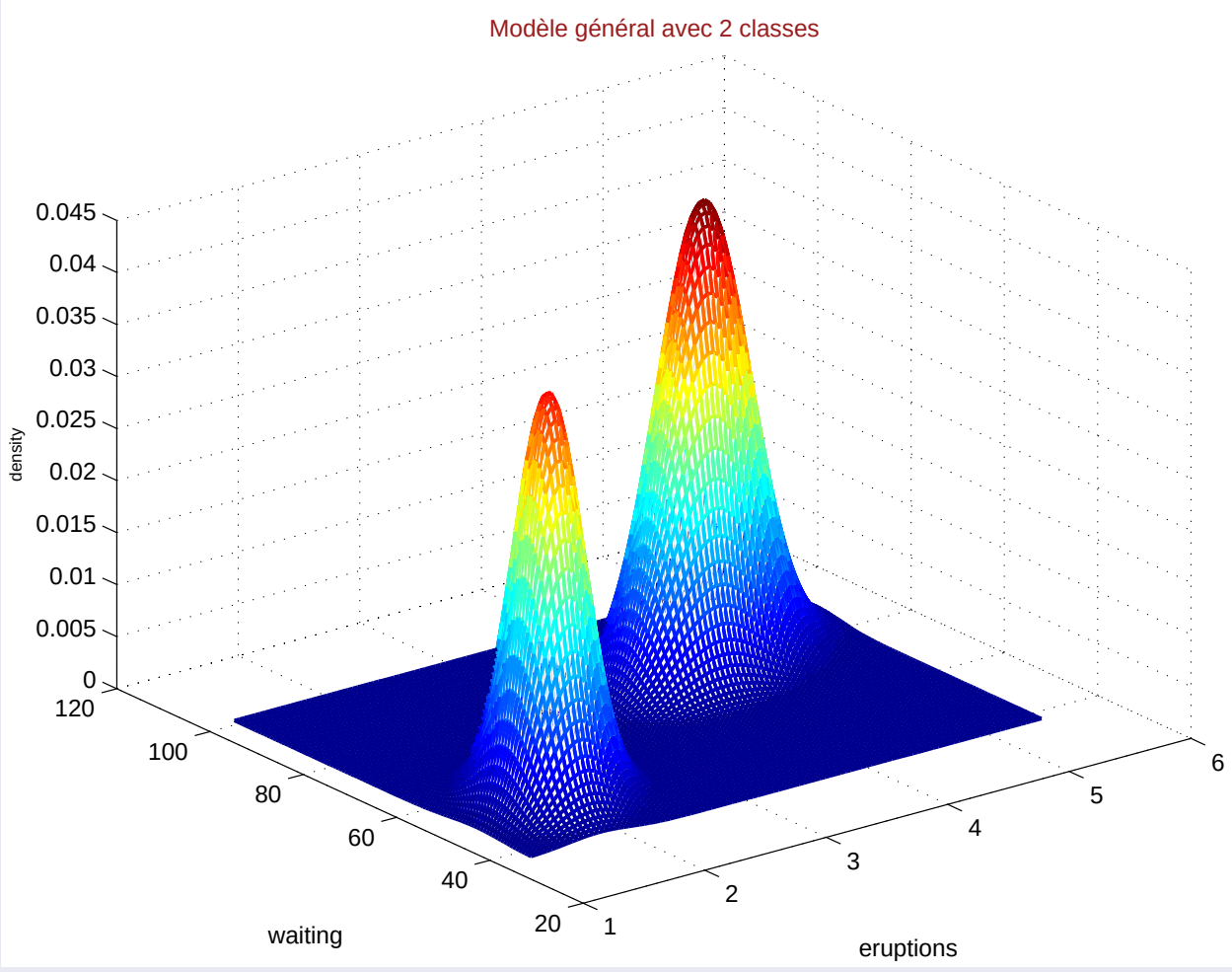
Visualisation des résultats



Visualisation des partitions obtenus par le MAP



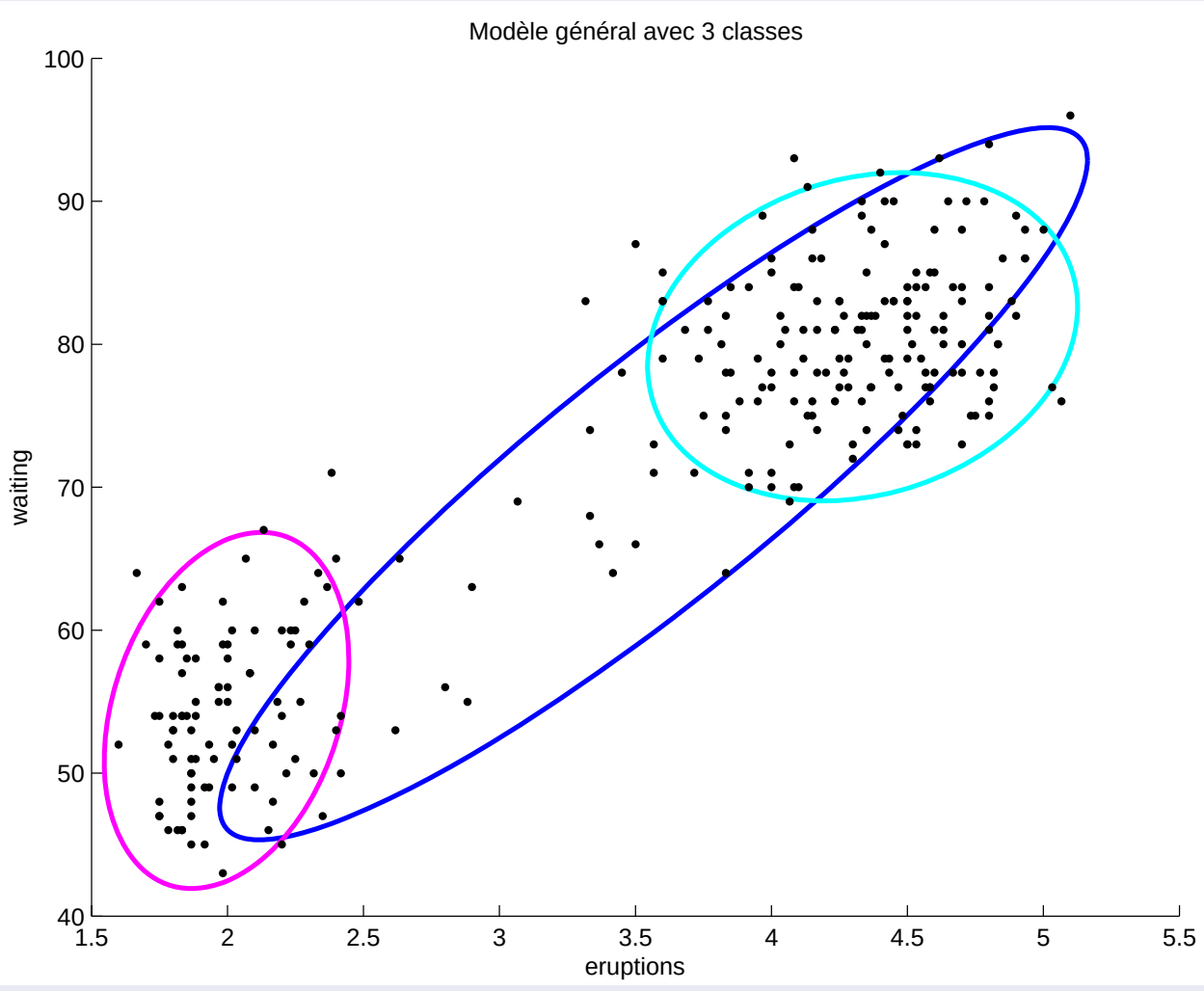
Visualisation de l'estimation de densité



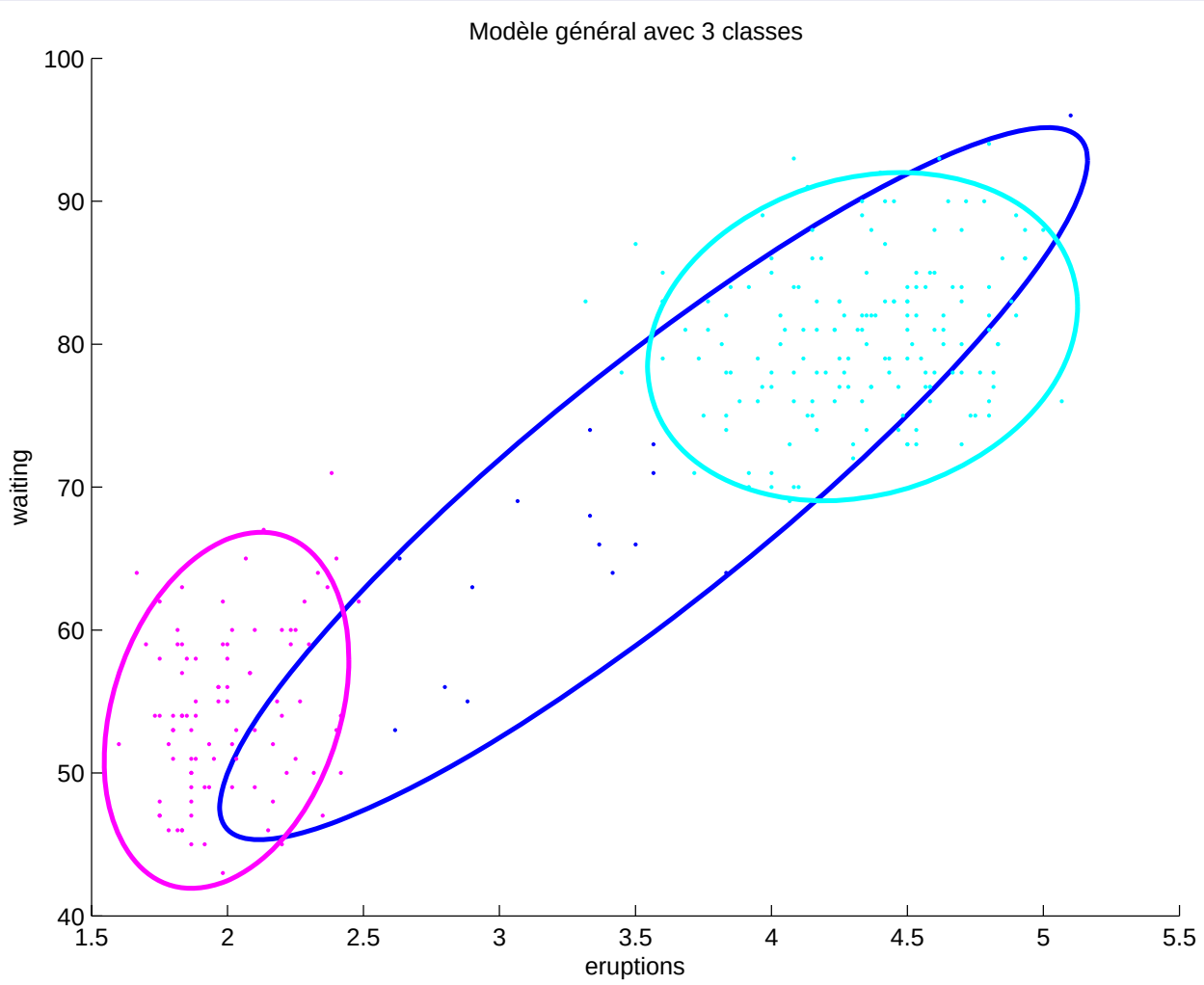
Utilisation du logiciel Mixmod avec 3 classes

```
 ----------------------------------------------------------------
*               Cluster size : list by cluster
 *                              .cluster 1    15
 *                              .cluster 2    92
 *                              .cluster 3    165
 *                 Proportions : list by cluster
                                .cluster 1     0.090213
                                .cluster 2     0.33275
                                .cluster 3     0.57704
 *                       Means : list by mean vector of clusters
                                .cluster 1
                                           3.56679       70.2397
                                .cluster 2
                                           1.99663       54.3831
                                .cluster 3
                                           4.33531       80.5227
 *                   Variances : list by clusters of variance matrix
                                .cluster 1
                                   0.5537     7.8501
                                    7.8501     134.8681
                                .cluster 2
                                   0.0439     0.3441
                                    0.3441     33.7411
                                .cluster 3
                                   0.1360     0.3588
                                    0.3588     28.5971
 *              Log-likelihood : -1119.214
  ----------------------------------------------------------------
```

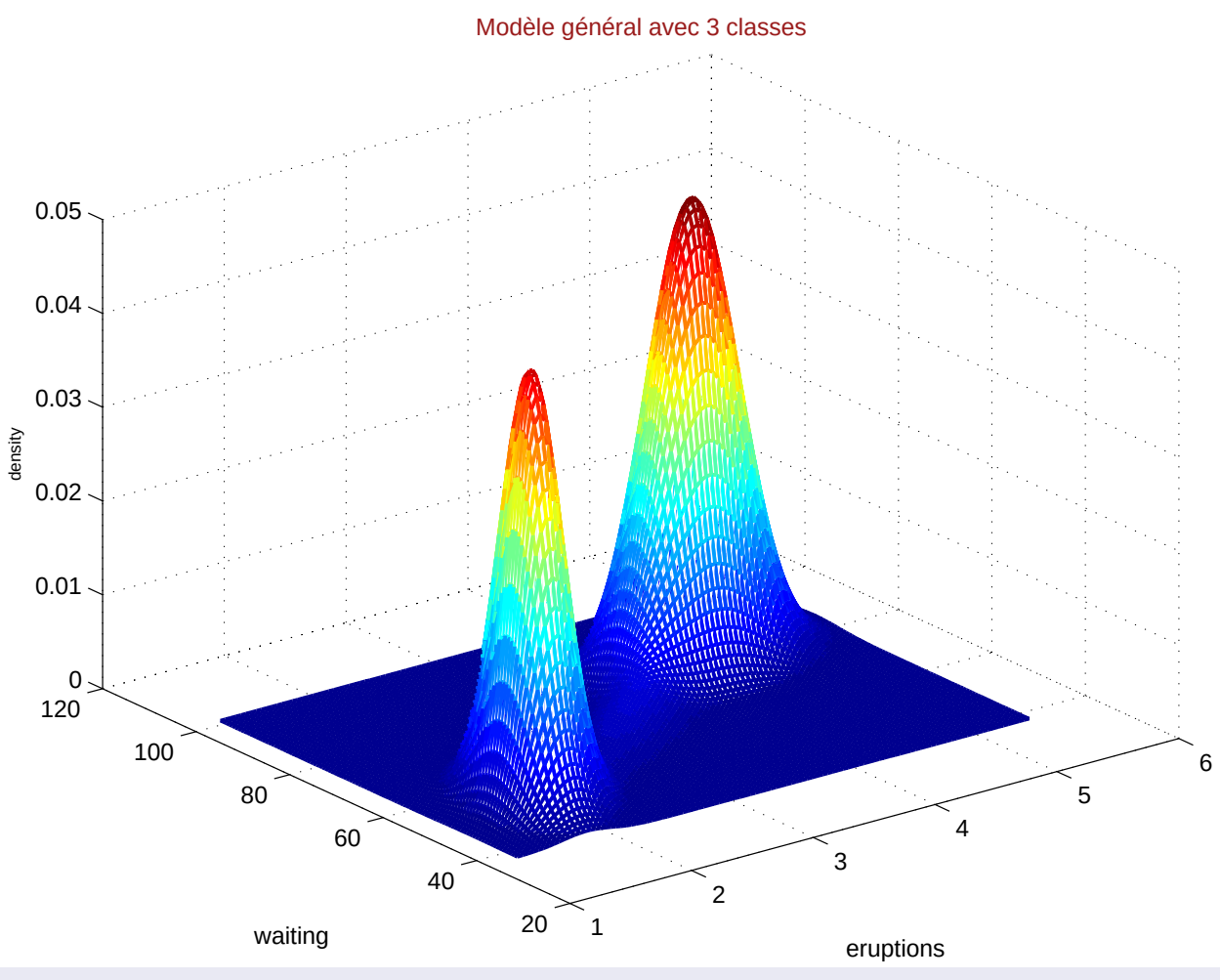
Visualisation des résultats



Visualisation des partitions obtenus par le MAP



Visualisation de l'estimation de densité



Nécessité

- Nombreux paramètres
- n petit, p ou K grand : trop de paramètres
- Nécessaire de diminuer le nombre de paramètres
- Modèles parcimonieux obtenus en imposant des contraintes
  - Proportions π_k
  - Matrices de variances Σ_k

Décomposition spectrale de la matrice de variance

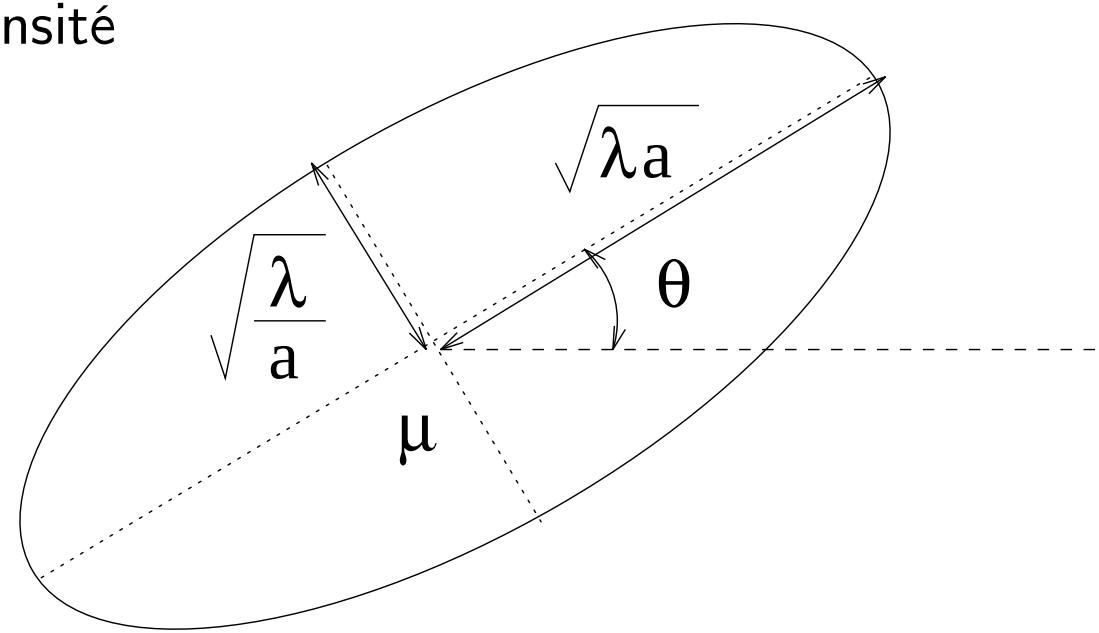
- Décomposition en valeurs propres et vecteurs propres $\Sigma_k = D_k B_k D_k'$
  - D_k matrice des vecteurs propres
  - B_k matrice diagonale des valeurs propres décroissantes (unicité)
- B_k décomposée en $B_k = \lambda_k A_k$ avec $|A_k| = 1$
- Finalement

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

  - A_k matrice diagonale de dét. 1 et à valeurs décroissantes : forme
  - D_k matrice orthogonale : orientation
  - λ_k réel positif : volume

Exemple dans $\mathbb{R}^2$

- D matrice de rotation définie par un angle θ
- A est une matrice diagonale de termes diagonaux a et $1/a$
- Ellipse d'équidensité



$$A = \begin{bmatrix} a & 0 \\ 0 & 1/a \end{bmatrix} \qquad D = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

Re-paramétrisation

- Le modèle de mélange est finalement paramétré par :
  - les proportions $\pi_1, \ldots, \pi_K$
  - les centres $\mu_1, \ldots, \mu_K$,
  - les volumes $\lambda_1, \ldots, \lambda_K$
  - les formes $A_1, \ldots, A_K$
  - les orientations $D_1, \ldots, D_K$
- Pour n petit ou d grand, il est alors possible de proposer des modèles plus parcimonieux en imposant des hypothèses restrictives variées et facilement interprétables :
  - Matrices de variance proportionnelles à la matrice identité
  - Matrices de variance identiques pour toutes les classes
  - ...
  - Aucune contrainte

28 modèles différents

- Famille générale : en supposant les proportions, volumes, orientations et formes identiques ou non suivant les classes, on obtient 16 différents modèles

$$[\pi_k \lambda_k D_k A_k], \quad \ldots \quad [\pi \lambda D A]$$

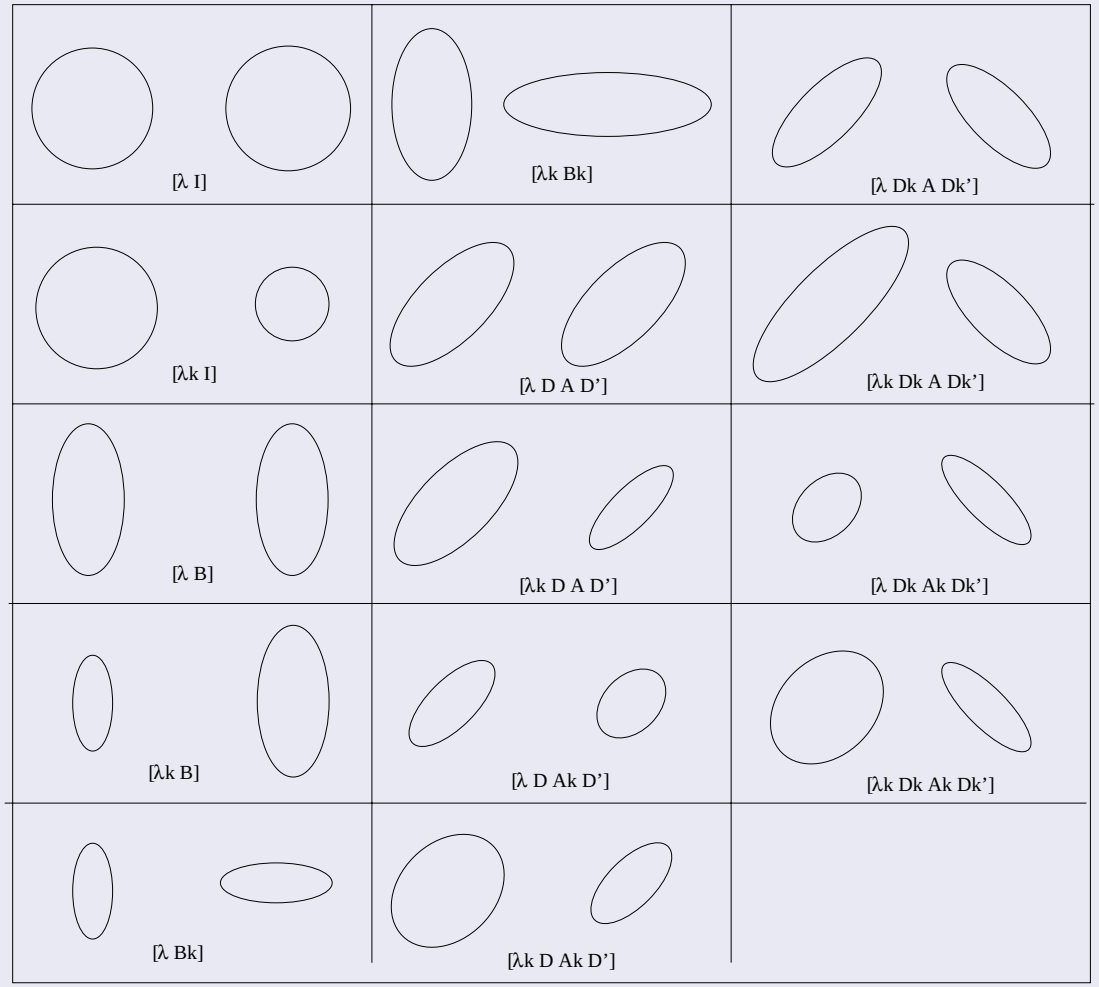
- Famille diagonale : en supposant de plus que les matrices de variances sont diagonales, on obtient 8 modèles

$$[\pi_k \lambda_k B_k], \quad \ldots \quad [\pi \lambda B]$$

- Famille sphérique : en supposant que la les matrices de variance sont proportionnelles à la matrice identité, on obtient 4 modèles

$$[\pi_k \lambda_k I] \quad [\pi \lambda_k I] \quad [\pi_k \lambda I] \quad [\pi \lambda I]$$

Illustration des 14 modèles différents



- Modification des algorithmes :
  - Forme particulière de la fonction d'affectation de l'étape C de l'algorithme CEM
  - Calcul des matrices de variance Σ_k de l'étape M
- Exemples étudiés :
  - Modèle $[\lambda I]$: forme sphérique et volumes identiques
  - Modèle $[\lambda_k I]$: forme sphérique et volumes différents
  - Modèle $[\lambda B]$: formes diagonales identiques
  - Modèle $[\Sigma]$: Formes identiques

Modèle $[\lambda I]$: forme sphérique et volumes identiques

- Classes ayant toutes la même distribution normale sphérique
- Les paramètres $\Sigma_1, \ldots, \Sigma_g$ se réduisent au réel λ
- Etape M :
  - Minimisation de $F(\lambda) = \frac{1}{\lambda}\text{trace}(S_W) + p \log \lambda$

$$\lambda = \frac{\text{trace}(S_W)}{p}$$

- Vrais.classifiante

$$L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{np}{2} \log \text{trace}(S_W) - \sum_k n_k \log \pi_k + A$$

Lien entre le modèle $[\lambda I]$ et l'algorithme des k-means

- Pour CEM et avec des proportions égales :
  - Etape d'affectation :
    - $d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log\pi_k$
    - Distance euclidienne habituelle $d^2(\boldsymbol{x}_i, \mu_k)$
  - Maximisation de de la vraisemblance classifiante
    - $L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{np}{2}\log \text{trace}(S_W) - \sum_k n_k \log \pi_k + A$
    - Minimisation du critère d'inertie intra-classe $\text{trace}(S_W)$
  - CEM : algorithme des centres mobiles *k-means*
- Utiliser le critère d'inertie revient à supposer que les classes sont sphériques, de même proportion et de même volume

Modèle $[\lambda_k I]$: forme sphérique et volumes différents

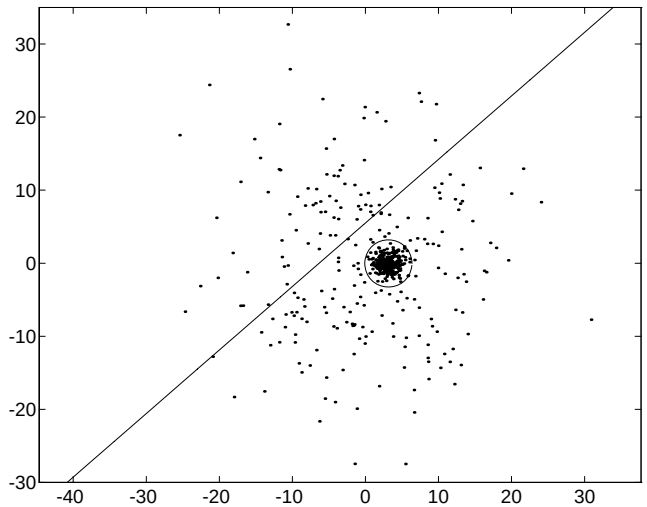
- Classes ayant toutes une distribution normale sphérique
- Les paramètres $\Sigma_1, \ldots, \Sigma_K$ se réduisent au vecteur $(\lambda_1, \ldots, \lambda_K)$
- Etape M : min. $F(\lambda_1, \ldots, \lambda_K) = \sum_k n_k \left(\frac{1}{\lambda_k} \text{trace}(S_k) + p \log \lambda_k \right)$:

$$\lambda_k = \frac{\text{trace}(S_k)}{p}$$

- Vraisemblance classifiante :
  $L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{p}{2} \sum_k n_k \log \text{trace}(S_k) - \sum_k n_k \log \pi_k + A$
- Pour CEM et avec des proportions égales :
  - Etape d'affectation :
    - $d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \log |\Sigma_k| - 2 \log \pi_k$
    - Distance $\frac{1}{\lambda_k} d^2(\boldsymbol{x}_i, \mu_k) + p \log(\lambda_k)$
  - Max. de la vraisemblance classifiante = minimisation de $\sum_k n_k \ln \text{trace}(S_k)$

Exemple

- Distance d'un point au centre modifiée par le volume de la classe
- Surfaces séparatrices : hyperplans $\longrightarrow$ hypersphères
- Ex. : 2 classes gaussiennes sphériques, même prop. et volumes très $\neq$



- Critère d'inertie intraclasse classique : la partition obtenue par la séparation avec la droite n'a aucun rapport avec la partition simulée
- Modèle à volume variable : la partition obtenue par la séparation avec le cercle est très proche de la classification simulée

Remarques

- Peu de différences entre les 2 modèles $[\lambda I]$ et $[\lambda_k I]$
- Modification pour prendre en compte des classes de volumes différents
- Augmentation du nombre de paramètres est assez faible $(K-1)$
- Les résultats peuvent être très différents
- Sans l'aide du modèle de mélange : difficile de proposer la distance

$$\frac{1}{\lambda_k} d^2(\boldsymbol{x}_i, \mu_k) + p \log(\lambda_k)$$

et le critère

$$\sum_k n_k \ln \text{trace}(S_k)$$

utilisés dans cette approche à partir d'une simple interprétation métrique

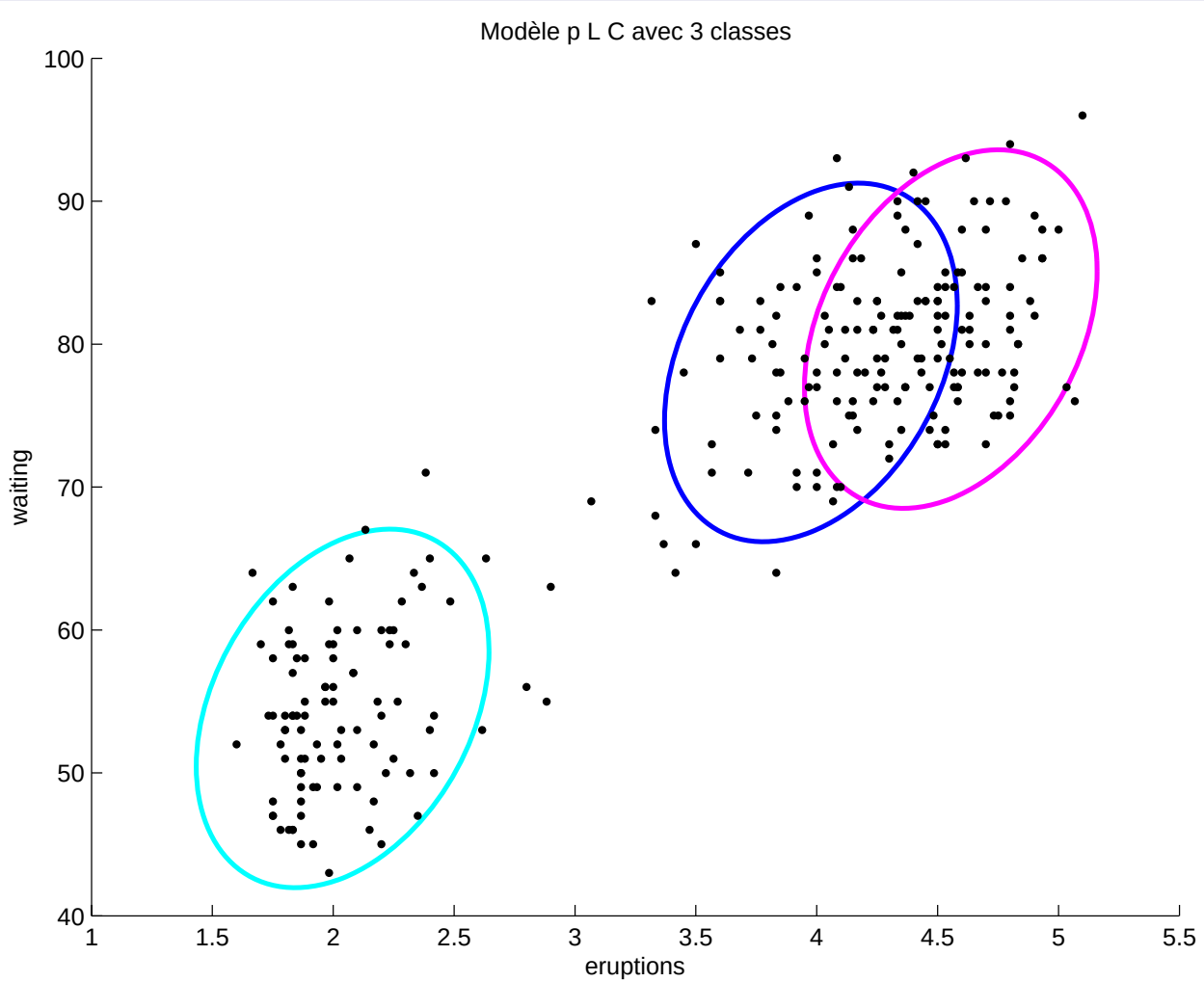
Modèle $[\lambda B]$: formes diagonales identiques

- La matrice de variance de chaque classe a maintenant la forme $\Sigma_k = \lambda B$ où B est une matrice diagonale de déterminant 1
- Simplification : $A = \lambda B$ où cette fois la matrice A est une matrice diagonale quelconque
- Les paramètres $\Sigma_1, \ldots, \Sigma_K$ se réduisent donc à la matrice A
- Etape M : min. $F(A) = n\left(\text{trace}(S_W A^{-1}) + \ln|A|\right)$: $A = \text{diag}(S_W)$
- Vrais.classifiante : $L_C(\mathbf{c}, \boldsymbol{\theta}) = -\frac{n}{2}\log|\text{diag}(S_W)| - \sum_k n_k \log \pi_k + C$
- Pour CEM et avec des proportions égales :
  - Etape d'affectation :
    - $d^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log \pi_k$
    - $d^2_{B^{-1}}(\mathbf{x}_i, \mu_k)$: distance euclidienne pondérée
  - Max. de la vraisemblance classifiante = minimisation de $|\text{diag}(S_W)|$

Modèle $[\lambda DAD']$: Formes identiques

- La matrice de variance de chaque classe a maintenant la forme $\Sigma_k = \Sigma$
- Les paramètres $\Sigma_1, \ldots, \Sigma_K$ se réduisent donc à la matrice Σ
- Etape M : min. $F(\Sigma) = n\left(\text{trace}(S_W\Sigma^{-1}) + \log|\Sigma|\right)$: $\Sigma = S_W$
- Vrais. classifiante :
  $L_C(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2}\left(np + n\ln|S_W|\right) - \sum_k n_k \log \pi_k + C$
- Pour CEM et avec des proportions égales :
  - Max. de la vraisemblance classifiante = minimisation de $|S_W|$

Modèle [$\lambda DAD'$]



Visualisation des partitions obtenus par le MAP

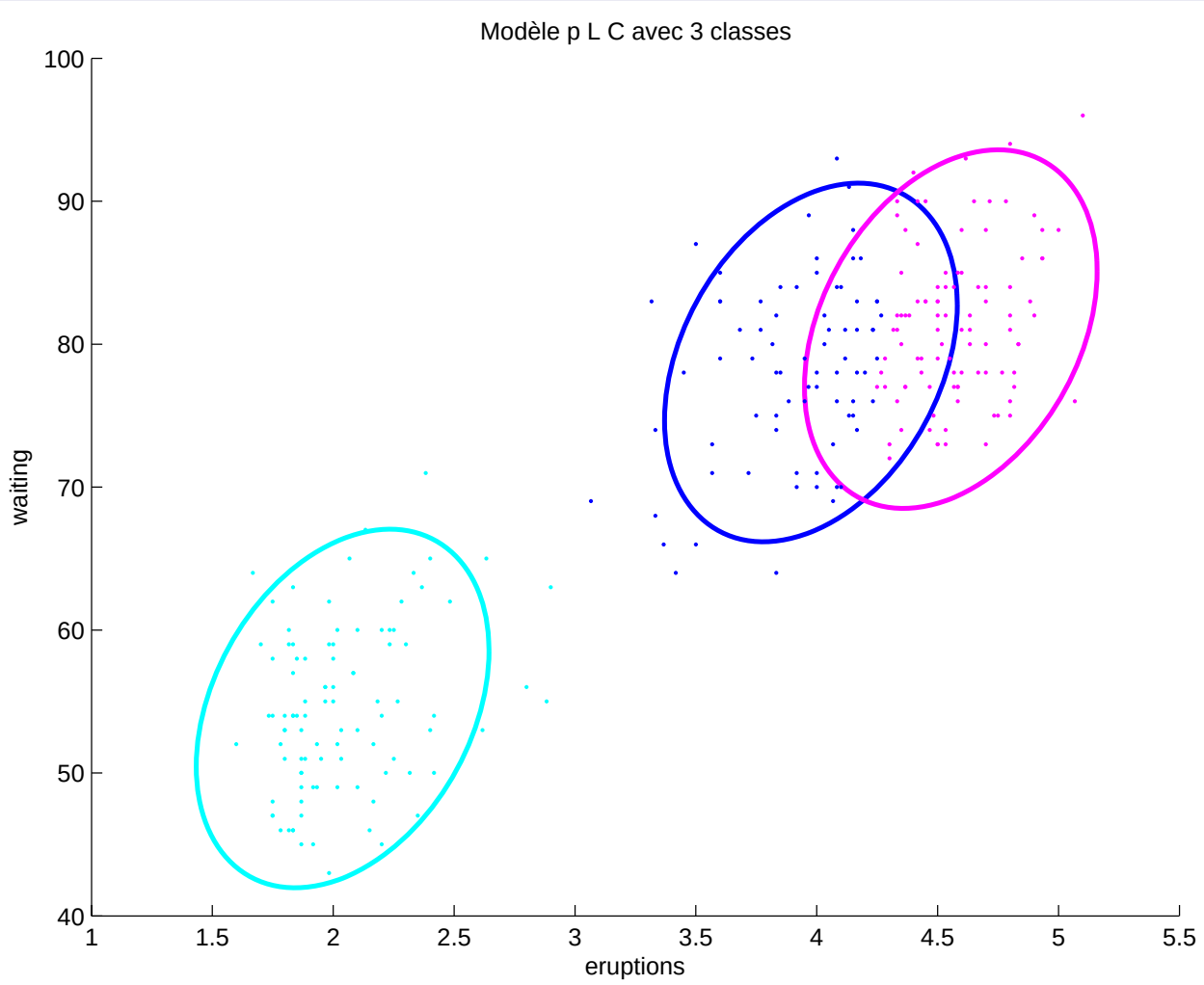


Tableau récapitulatif

modèle	nombre de paramètres	étape M	critère pour CEM (π)
$[\lambda DAD']$	$\alpha+\beta$	Exp.	$\lvert S_W\rvert$
$[\lambda_k DAD']$	$\alpha+\beta+K-1$	PI	-
$[\lambda DA_k D']$	$\alpha+\beta+(K-1)(d-1)$	PI	-
$[\lambda_k DA_k D']$	$\alpha+\beta+(K-1)d$	PI	-
$[\lambda D_k AD'_k]$	$\alpha+K\beta-(K-1)d$	Exp.	$\lvert\sum_k n_k\Omega_k\rvert$
$[\lambda_k D_k AD'_k]$	$\alpha+K\beta-(K-1)(d-1)$	PI	-
$[\lambda D_k A_k D'_k]$	$\alpha+K\beta-(K-1)$	Exp.	$\sum_k n_k\lvert S_k\rvert^{\frac{1}{d}}$
$[\lambda_k D_k A_k D'_k]$	$\alpha+K\beta$	Exp.	$\sum_k n_k \ln\lvert S_k\rvert$
$[\lambda B]$	$\alpha+d$	Exp.	$\lvert\mathrm{diag}(S_W)\rvert$
$[\lambda_k B]$	$\alpha+d+K-1$	PI	-
$[\lambda B_k]$	$\alpha+Kd-K+1$	Exp.	$\sum_k n_k\lvert\mathrm{diag}(S_k)\rvert^{\frac{1}{d}}$
$[\lambda_k B_k]$	$\alpha+Kd$	Exp.	$\Sigma_k n_k \ln\mathrm{diag}(S_k)$
$[\lambda I]$	$\alpha+1$	Exp.	$\mathrm{trace}(S_W)$
$[\lambda_k I]$	$\alpha+d$	Exp.	$\sum_k n_k \ln\mathrm{trace}(S_k)$

Exp : explicite, PI : procédure itérative

CEM et les algorithmes de classification classiques

modèle	distance	critère	remarques
$\pi, \lambda I$	$d^2(x_i, \mu_k)$	$\text{trace}(S_W)$	k-means
$\pi, \lambda_k I$	$\frac{d^2(x_i,\mu_k)}{\lambda_k} + d \ln(\lambda_k)$	$\sum_k n_k \ln \text{tr}(S_k)$	
$\pi, \lambda B$	$d^2_{B^{-1}}(x_i, \mu_k)$	$\text{diag}(S_W)$	class. + pondérations
π, Σ	$d^2_{\Sigma^{-1}}(x_i, \mu_k)$	$\lvert S_W \rvert$	Friedman et Rubin
π, Σ_k	$d^2_{\Sigma_k^{-1}}(x_i, \mu_k)$	$\sum_k n_k \ln \lvert S_k \rvert$	Scott et Symons

Remarques

- Propriétés d'invariance
  - Famille générale : résultats invariants pour toute transformation linéaire des données
  - Famille diagonale : résultats invariants pour toute normalisation des variables
  - Famille sphérique : résultats invariants pour toute transformation isométrique
- Modèles sphériques et diagonaux : hypothèse des « classes latentes »qui suppose que les variables initiales sont indépendantes conditionnellement à la connaissance du composant

Cinquième partie V

Mélange multinomial multivarié

Les données qualitatives

- Les n observations à classifier sont décrits par d variables qualitatives
- Chaque variable j a m_j niveaux de réponse (modalités)

Les données $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ sont définies par

$$\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$$

où

$$\begin{cases} x_i^{jh} = 1 & \text{si } i \text{ prend le niveau } h \text{ pour la variable } j \\ x_i^{jh} = 0 & \text{sinon.} \end{cases}$$

$\mathbf{x}$ peut être vu comme une matrice de dimension (n, m) où $m = \sum_{j=1}^{d} m_j$ est le nombre total de modalités

Le modèle des classes latentes

Les données sont supposées provenir d'un mélange de g distributions multivariées multinomiales de densité

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k f_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \sum_k \pi_k \prod_{j,h} (\alpha_k^{jh})^{x_i^{jh}}$$

où $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \alpha_1^{11}, \ldots, \alpha_g^{dm_d})$ est le paramètre du modèle des classes latentes à estimer :

- α_k^{jh} : probabilité que la variable j prenne le niveau h dans la classe k
- π_k : proportions du mélange

Le modèle des classes latentes suppose que les variables sont conditionnellement indépendantes connaissant la classe

Étape E

- Les étapes E des algorithmes EM et CEM sont identiques
- Pas de problème particulier
- Calcul des des probabilités d'appartenance des $\mathbf{x}_i$ aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{(c)})}{\sum_\ell \pi_\ell^{(c)} \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_\ell^{(c)})}$$

Étape de classification C

- Etape supplémentaire de classification pour CEM
- La partition $z^{(c+1)}$ est obtenue en rangeant chaque x_i dans la classe maximisant $t_{ik}^{(c)}$:

$$z_{ik}^{(c+1)} = \begin{cases} 1 & \text{si } k = \operatorname{argmax}_k \ t_{ik}^{(c)} \\ 0 & \text{sinon} \end{cases}$$

- Chaque x_i est donc rangé dans la classe qui
  - Maximise $\pi_k \varphi(x_i; \alpha_k)$
  - Maximise $\log(\pi_k \varphi(x_i; \alpha_k))$ ou encore

$$\sum_{j,h} x_i^{jh} \log \alpha_k^{jh} + \log \pi_k$$

Notations

- $\mathbf{c} = (c_{ik})$ avec
  - $c_{ik} = t_{ik}^{(c)}$ pour EM (partition floue)
  - $c_{ik} = z_{ik}^{(c)}$ pour CEM (partition)
- $n_k = \sum_i c_{ik}$
- $n_k^{jh} = \sum_i c_{ik} x_i^{jh}$
- Lorsque $\mathbf{c}$ correspond à une partition
  - n_k est le cardinal de la classe k
  - n_k^{jh} est le nombre de fois où la modalité h de la variable j a été prise dans la classe k

Problème posé

Maximisation en $\boldsymbol{\theta}$ de

$$L_C(\theta, \mathbf{c}) = \sum_{i,k} c_{ik} \log \left[\pi_k \varphi(\boldsymbol{x}_i | \boldsymbol{\alpha}_k)\right]$$

qui s'écrit pour le modèle de mélange multinomial

$$\sum_k \sum_{j,h} n_k^{jh} \log \alpha_k^{jh} + \sum_k n_k \log \pi_k$$

Résultat

- $\forall k, j$: maximisation de $\sum_h n_k^{jh} \log \alpha_k^{jh}$ sous la contrainte $\sum_h \alpha_{jh} = 1$:

$$\alpha_k^{jh} = \frac{n_k^{jh}}{n_k}$$

- $\forall k$: maximisation de $\sum_k n_k \log \pi_k$ sous la contrainte $\sum_k \pi_k = 1$:

$$\pi_k = n_k / n$$

- La valeur de L_c maximisée

$$L_C(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{k,j,h} n_k^{jh} \log n_k^{jh} - \sum_k n_k \log n_k + \sum_k n_k \log \pi_k$$

Diminution du nombre de paramètres

- Nombre de paramètres du modèle de classes latentes

$$(K-1)+K*\sum_{j}(m_j-1)$$

- Beaucoup plus petit que le nombre de paramètres du modèle log-linéaire complet

$$\prod_{j=1}^{d} m_j$$

- Exemple : $K=5$, $d=10$ et $m_j=4$ pour toutes les variables : 154 et 10^4 paramètres
- Toutefois, nombre de paramètres pouvant encore trop grand : nécessité de disposer de modèles plus parcimonieux

Interprétabilité des résultats

- Chaque classe est caractérisée par les $\alpha_k^{j\ell}$
- Interprétation difficile
- Dans les modèles parcimonieux proposés, chaque classe est caractérisée
  - par un vecteur de variables qualitatives (forme identique aux données initiales)
  - par un élément (vecteur ou réel) de dispersion

Re-paramétrisation

- Pour chaque composant k et chaque variable j :

$$(\alpha_k^{j1}, \ldots, \alpha_k^{jm_j}) \longrightarrow (a_k^{j1}, \ldots, a_k^{jm_j}, \varepsilon_k^{j1}, \ldots, \varepsilon_k^{jm_j})$$

  - le vecteur binaire $a_k^{j1}, \ldots, a_k^{jm_j}$ fournit la modalité la plus probable

$$a_k^{jh} = \begin{cases} 1 & \text{si } h = \arg\max_h \alpha_k^{jh} \\ 0 & \text{sinon} \end{cases}$$

  - les ε_k^{jh} peuvent être vus comme des valeurs de dispersion :

$$\varepsilon_k^{jh} = \begin{cases} 1 - \alpha_k^{jh} & \text{si } a_k^{jh} = 1 \\ \alpha_k^{jh} & \text{si } a_k^{jh} = 0 \end{cases}$$

- Exemple : $(0.7, 0.2, 0.1) \longrightarrow (1, 0, 0, \quad 0.3, 0.2, 0.1)$

Modèle $[\varepsilon_k^j]$ (1)

- Contraintes sur les paramètres de dispersion $\boldsymbol{\varepsilon}_k^j = (\varepsilon_k^{j1}, \ldots, \varepsilon_k^{jm_j})$
- $\boldsymbol{\varepsilon}_k^j$ caractérisé par une seule valeur réelle :

$$(a, \ldots, a, b, a, \ldots, a) \quad \text{avec} \quad 1 - b > a$$

- $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ prend alors la forme

$$(\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j) \quad \text{avec} \quad \gamma_k^j > 1/m_j$$

Modèle $[\varepsilon_k^j]$ (2)

- $\boldsymbol{\alpha}_k^j$ peut être ainsi décomposé suivant les 2 paramètres :
  - $\mathbf{a}_k^j = (a_k^{j1}, \ldots, a_k^{jm_j})$:
    - $a_k^{jh} = 1$ si h correspond au rang de γ_k^j
    - sinon
  - $\varepsilon_k^j = 1 - \gamma_k^j$: probabilité que, pour le composant k, la variable j ne prenne pas la valeur majoritaire
- Reparamétrisation de chaque distribution multinomiale par un centre $\mathbf{a}_k^j$ et une dispersion ε_k^j
- Interprétation similaire à la distribution gaussienne paramétrée par un centre et une variance
- Exemple : $\alpha_k^j = (0.7, 0.15, 0.15) \longrightarrow a_k^j = 1$ et $\varepsilon_k^j = 0.3$

Modèle $[\varepsilon_k^j]$ (3)

- Relation entre ce nouveau paramétrage et le paramétrage initial :

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j & \text{si } h = h(k,j) \\ \dfrac{\varepsilon_k^j}{m_j - 1} & \text{sinon} \end{cases}$$

- La densité φ peut être réécrite avec $\mathbf{a}_k = (\mathbf{a}_k^j; j = 1, \ldots, d)$ et $\boldsymbol{\varepsilon}_k = (\varepsilon_k^j; j = 1, \ldots, d)$

$$\varphi(\boldsymbol{x}_i|\boldsymbol{\alpha}_k) = \varphi(\boldsymbol{x}_i|\mathbf{a}_k, \boldsymbol{\varepsilon}_k) = \prod_{j,h} \left((1 - \varepsilon_k^j)^{a_k^{jh}} \left(\frac{\varepsilon_k^j}{m_j - 1} \right)^{1 - a_k^{jh}} \right)^{x_i^{jh}}$$

Cinq modèles

En notant $[\varepsilon_k^{jh}]$ le modèle initial et en imposant de nouvelles contraintes, on obtient finalement 5 modèles multinomiaux :

- $[\varepsilon_k^{jh}]$ (le modèle des classes latentes standard) : la dispersion dépend du composant, de la variable et de la modalité
- $[\varepsilon_k^{j}]$: la dispersion dépend du composant et de la variable mais pas de la modalité
- $[\varepsilon_k]$: la dispersion ne dépend que du composant
- $[\varepsilon^{j}]$: la dispersion dépend de la variable mais pas du composant ni de la modalité
- $[\varepsilon]$: la dispersion ne dépend ni du composant, ni de la variable, ni de la modalité

Nombre de paramètres des cinq modèles

modèle	nombre de paramètres
$[\varepsilon]$	$\delta + 1$
$[\varepsilon^j]$	$\delta + d$
$[\varepsilon_k]$	$\delta + g$
$[\varepsilon_k^j]$	$\delta + gd$
$[\varepsilon_k^{jh}]$	$\delta + g\sum_{j=1}^{d}(m_j - 1)$

$\delta = K - 1$ dans le cas des proportion libres et $\delta = 0$ dans le cas des proportions égales

Calcul des centres et des proportions

- Seule l'étape M a besoin d'être détaillée
- Objectif : maximiser en $\boldsymbol{\theta}$ la log-vraisemblance classifiante

$$L_C(\mathbf{c}, \boldsymbol{\theta}) = \sum_{k,j} \left(\log \frac{(m_j - 1)(1 - \varepsilon_k^j)}{\varepsilon_k^j} \sum_h n_k^{jh} a_k^{jh} + n_k \log \varepsilon_k^j \right) + \sum_k n_k \log \pi_k + A$$

- Proportions (si elles sont libres) :

$$\pi_k = \frac{n_k}{n}$$

- Centres :
  - $1 - \varepsilon_k^j > \varepsilon_k^j \quad \forall k, j \Longrightarrow$ les a_k^{jh} doivent donc maximiser $\sum_h a_k^{jh} n_k^{jh}$
  - k^j est la modalité majoritaire de la classe k pour chaque variable j

Calcul des termes de dispersion (1)

- Notations :
  - $e_k^j = n_k - m_k^{jh}$: nb de désaccords avec la modalité majoritaire pour la classe k et la variable j
  - $e^j = \sum_k e_k^j$: nb de désaccords avec la modalité majoritaire pour la variable j et pour toutes les classes
  - $e_k = \sum_j e_k^j$: nb de désaccords avec la modalité majoritaire pour la classe k et toutes les variables
  - $e = \sum_{k,j} e_k^j$: nb de désaccords avec la modalité majoritaire pour toutes les classes et toutes les variables
- L_C s'écrit alors, à une cste près,

$$L_C(\mathbf{c}, \boldsymbol{\theta}) = \sum_{k,j} \left((n_k - e_k^j) \log(1 - \varepsilon_k^j) + e_k^j \log(\varepsilon_k^j) \right) + n_k \log \pi_k$$

Calcul des termes de dispersion (1)

- On obtient à chaque fois les proportions de désaccord associées :
  - Modèle $[\varepsilon_k^j]$:
$$\varepsilon_k^j = e_k^j / n_k \quad \forall j, k$$
  - Modèle $[\varepsilon^j]$
$$\varepsilon^j = e^j / n \quad \forall j$$
  - Modèle $[\varepsilon_k]$
$$\varepsilon_k = e_k / (n_k d) \quad \forall k$$
  - Modèle $[\varepsilon]$
$$\varepsilon = e / (nd)$$

Remarque

- Modèle $[\varepsilon]$ et proportions égales
- Vraisemblance classifiante :

$$L_C(\boldsymbol{\theta}, \mathbf{c}) = \log(\frac{\varepsilon}{1-\varepsilon}) \sum_{i,k} c_{ik} d(\boldsymbol{x}_i, \mathbf{a}_k) + nd \log(1-\varepsilon)$$

où $d(\boldsymbol{x}_i, \mathbf{a}_k)$ est une distance mesurant le nombre de modalités différentes entre le vecteur $\boldsymbol{x}_i$ et le centre $\mathbf{a}_k$

- Etape de classification de CEM : affectation de l'individus i à la classe k qui minimise $d(\boldsymbol{x}_i, \mathbf{a}_k)$
- Etape M : coordonnées a_k^j des centres $\mathbf{a}_k$ sont obtenus en prenant les modalités majoritaires

Les données de Stouffer

216 individus mesurés avec 4 variables binaires

$S1$	$S2$	$S3$	$S4$	fréquences
1	1	1	1	42
1	1	1	0	23
1	1	0	1	6
1	1	0	0	25
1	0	1	1	6
1	0	1	0	24
1	0	0	1	7
1	0	0	0	38
0	1	1	1	1
0	1	1	0	4
0	1	0	1	1
0	1	0	0	6
0	0	1	1	2
0	0	1	0	9
0	0	0	1	2
0	0	0	0	20

Les résultats

- Modèles utilisés :
  - Modèle des classes latentes : 9 paramètres
  - Modèle log-linéaire avec interaction d'ordre 2 : 11 paramètres
- Déviances :
  - Modèle log-linéaire : 7.11
  - Modèle de classes latentes : 2.72
- Paramètres pour le modèle de classes latentes :

Classe	p_k	a_k^{11}	a_k^{21}	a_k^{12}	a_k^{22}
1	0.279	0.993	0.940	0.927	0.769
2	0.721	0.714	0.330	0.354	0.132