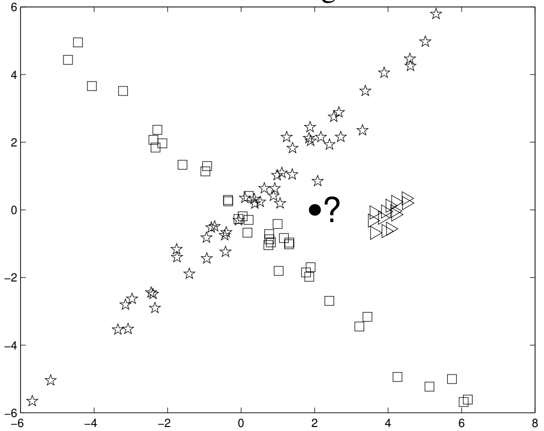
# Classification et discrimination

**Christophe Biernacki**
UMR CNRS 6623 – Université de Franche-Comté

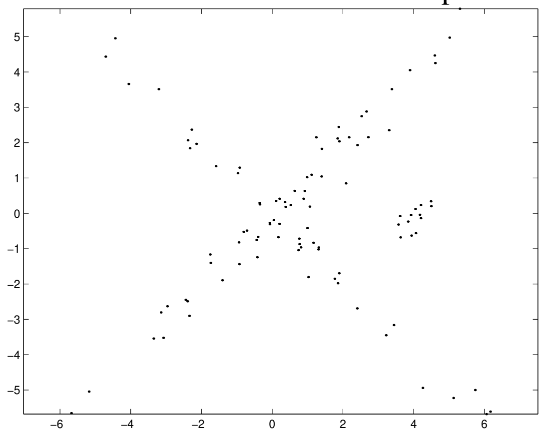
Journée didactique IS2 sur les mélanges de lois de probabilité
9 mars 2001

# Les problèmes

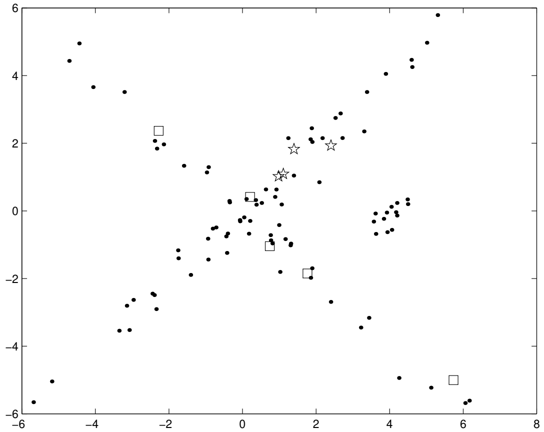
discrimination : estimer une règle de classement




classification : estimer une partition



classification avec connaissance partielle : estimer le reste de la partition



## PLAN

**La discrimination**

**La classification automatique**

**La classification avec connaissance partielle**

**La discrimination généralisée**

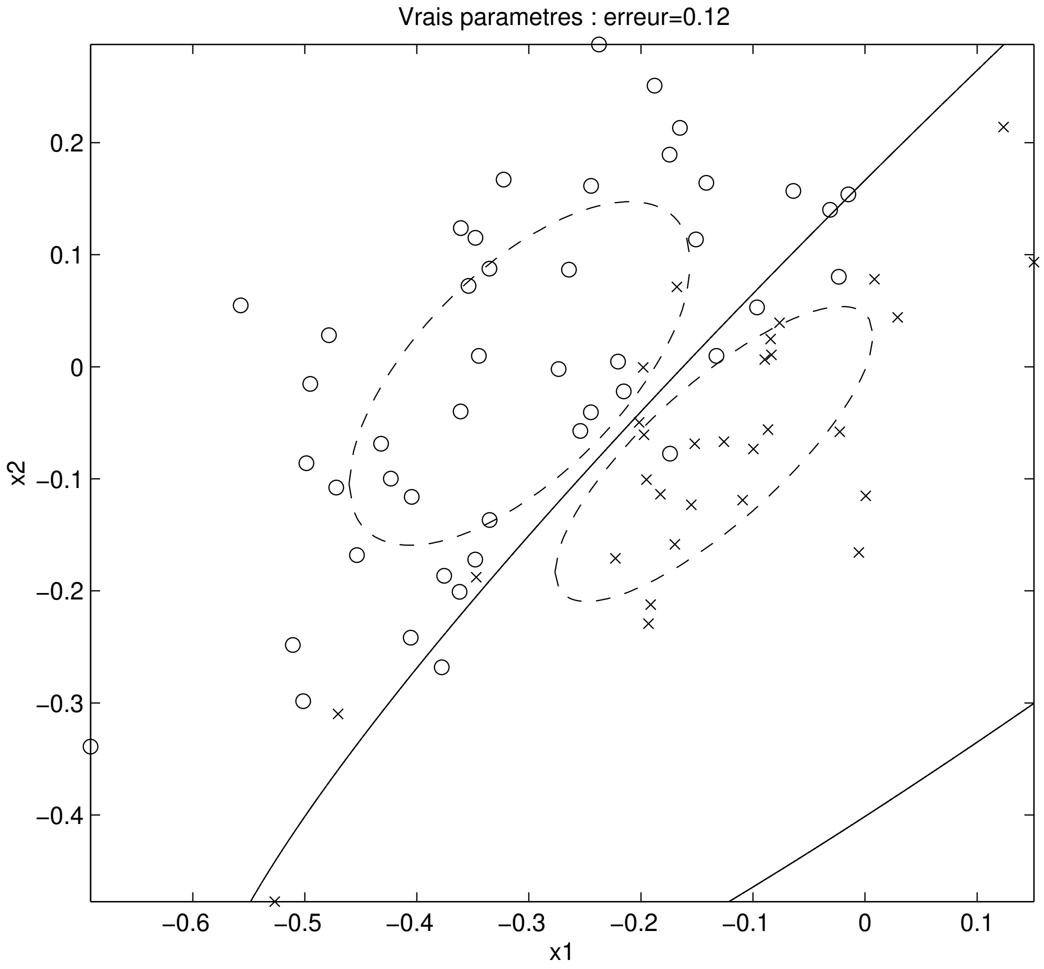
**Conclusion**

## Domaines d'application variés

- ***médecine*** : diagnostic d'un patient à partir d'examens cliniques
- ***industrie*** : état d'un processus industriel à partir de mesures
- ***banques*** : décision d'accord de crédits en fonction du profil client
- ***météorologie*** : évaluation du risque d'avalanches à partir de variables liées à l'atmosphère et à la neige
- ...

# Un exemple en médecine

30 femmes normales et 45 femmes hémophiles A mesurées par le log(activité AHF) et le log(antigène AHF)



## Les données

- $n$ ***observations*** $x = (x_1, \ldots, x_n)$ et $n$ ***labels*** $z = (z_1, \ldots, z_n)$
- nature de $x_i$ : quantitatif continu (longueur, etc), quantitatif discret (nombre de défauts, etc), qualitatif (réponse oui/non, etc)
- nature de $z_i$ : numéro du groupe de $x_i$ parmi $K$ possibles

## L'hypothèse de mélange

$$(x_1, z_1), \ldots, (x_n, z_n) \overset{i.i.d.}{\sim} (X, Z)$$

- groupe $k$ modélisé par une loi de paramètre $a_k$ : $X_{|Z=k} \sim \phi(\cdot, a_k)$
- groupe $k$ a une probabilité $p_k$ : $P(Z = k) = p_k$
- paramètre du mélange : $\theta = (p_1, \ldots, p_K, a_1, \ldots, a_K)$

## Classement d'une nouvelle observation $y$

1. estimateur du maximum de vraisemblance $\hat{\theta}$

$$L(\theta) = \sum_{k=1}^{K} \sum_{i/z_i=k} \ln(p_k \phi(x_i|a_k))$$

2. probabilité que $y$ provienne du groupe $k$

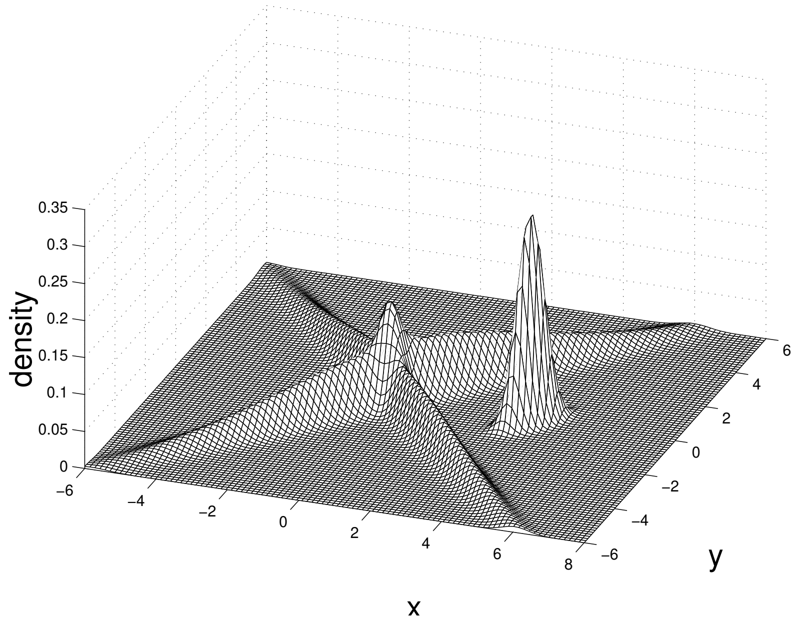
$$t_k(y|\hat{\theta}) = \frac{\hat{p}_k \phi(y|\hat{a}_k)}{\sum_{\ell=1}^{K} \hat{p}_\ell \phi(y|\hat{a}_\ell)}$$

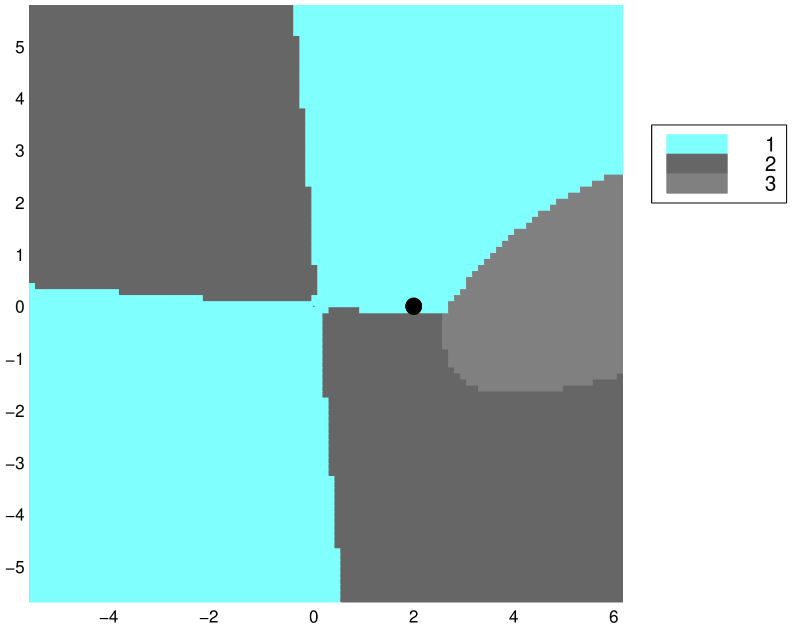
3. $y$ est affecté à la classe la plus probable (MAP)

# Mélange gaussien le plus général

$$\phi(x|a_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)\right)$$

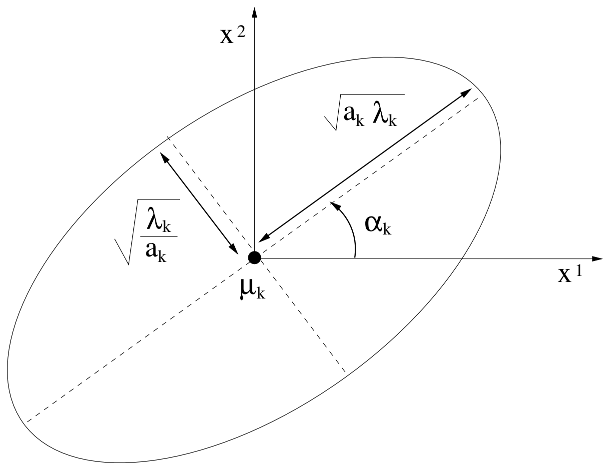
$x$ : $d$ valeurs réelles, $\mu_k$ : centre, $\Sigma_k$ : matrice de dispersion groupe $k$



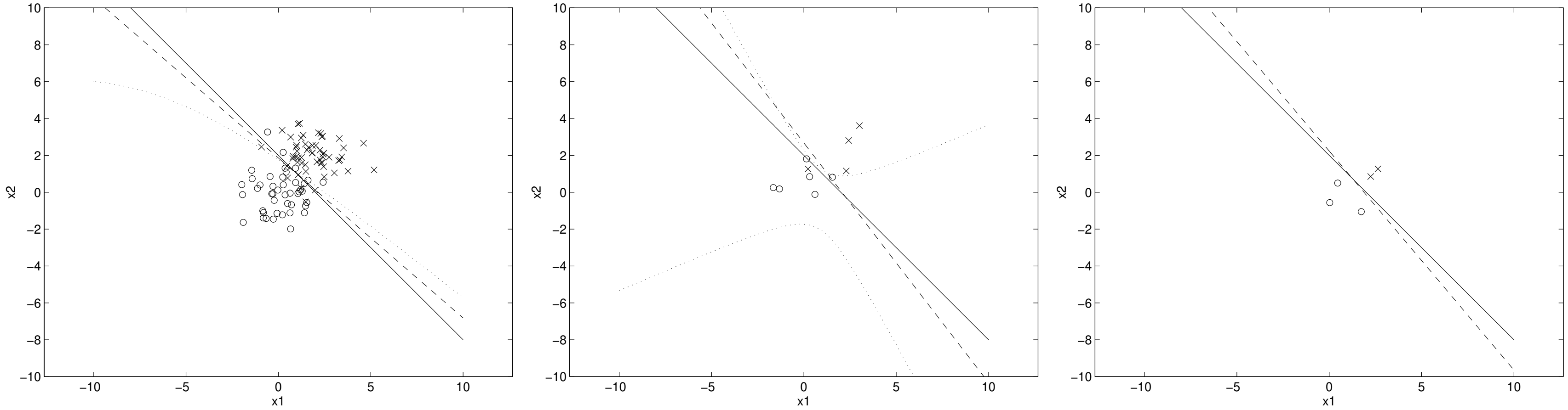
# 14 modèles gaussiens

$$\Sigma_k = \underbrace{\lambda_k}_{volume} \cdot \underbrace{D_k}_{orientation} \cdot \underbrace{A_k}_{forme} \cdot D_k'$$




$$A_k = \begin{pmatrix} a_k & 0 \\ 0 & 1/a_k \end{pmatrix}, \quad D_k = \begin{pmatrix} \cos\alpha_k & -\sin\alpha_k \\ \sin\alpha_k & \cos\alpha_k \end{pmatrix}$$

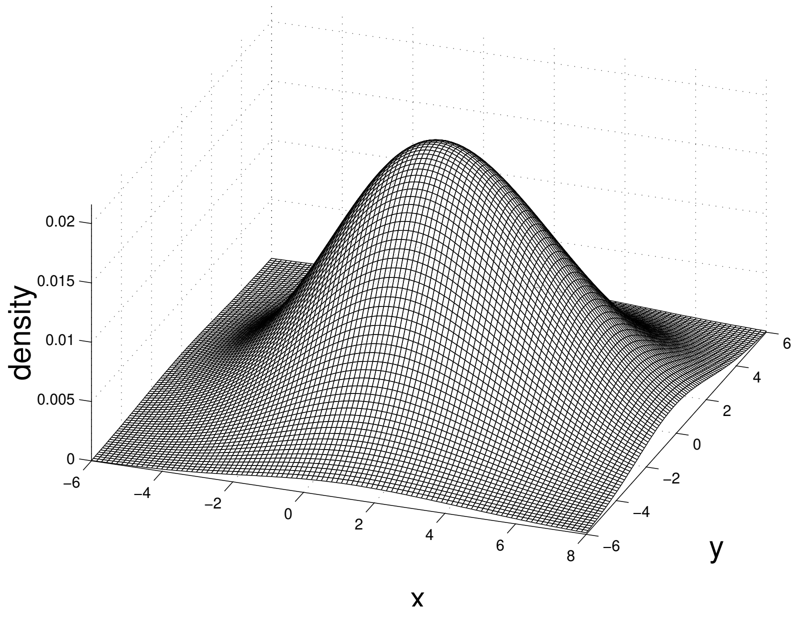
# Intérêt d'un modèle simple

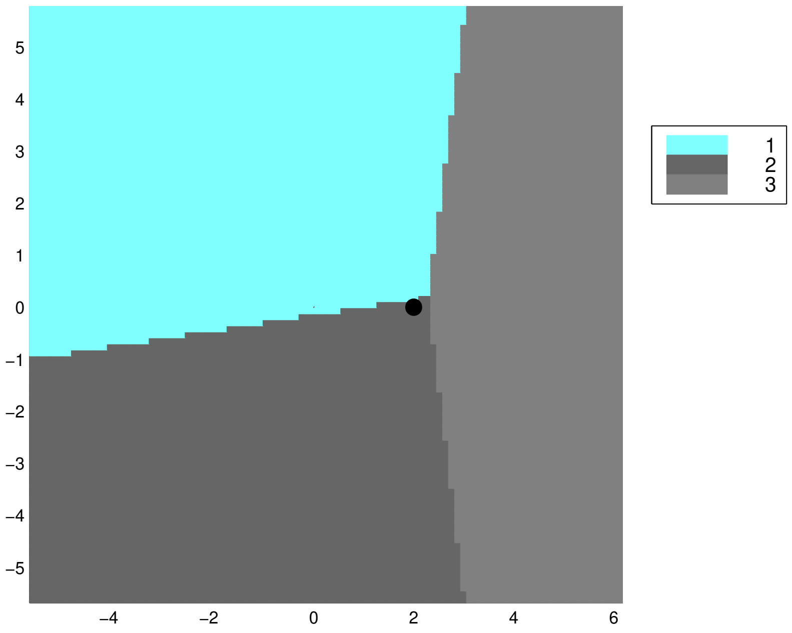



$n = 100$ $\qquad$ $n = 10$ $\qquad$ $n = 5$

—— vraie frontière $\qquad$ - - - frontière sphérique $\qquad$ . . . frontière quadratique

# Inconvénient d'un modèle trop simple



# Critères de choix de modèles

- ***utilisateur*** : connaissances *a priori*
- ***validation croisée*** : estimation de l'erreur de classement
- ***critères BIC, AIC*** : pénalisation de la vraisemblance $L - f(taille\ \theta)$

## Exemple

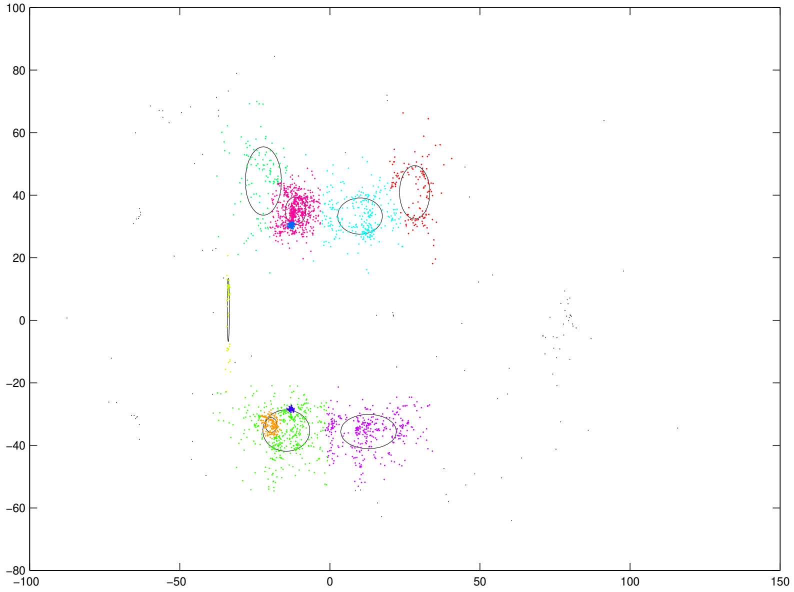
| critère | mod. le plus simple | mod. le plus complexe |
|---|---|---|
| VC | 0.52 | ***0.03*** |
| BIC | 1173.6 | ***777.6*** |

# La classification automatique

## Données : $x$

## Exemple industriel

Recherche de défauts dans une cuve



# Estimation des labels $z$

1. $\hat{\theta}$ : deux approches
   - ***mélange*** : estimateur du maximum de vraisemblance L

$$L(\theta) = \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} p_k \phi(x_i | a_k)) \right)$$

   - ***classification*** : estimateur du MV complétée CL (défauts théoriques)

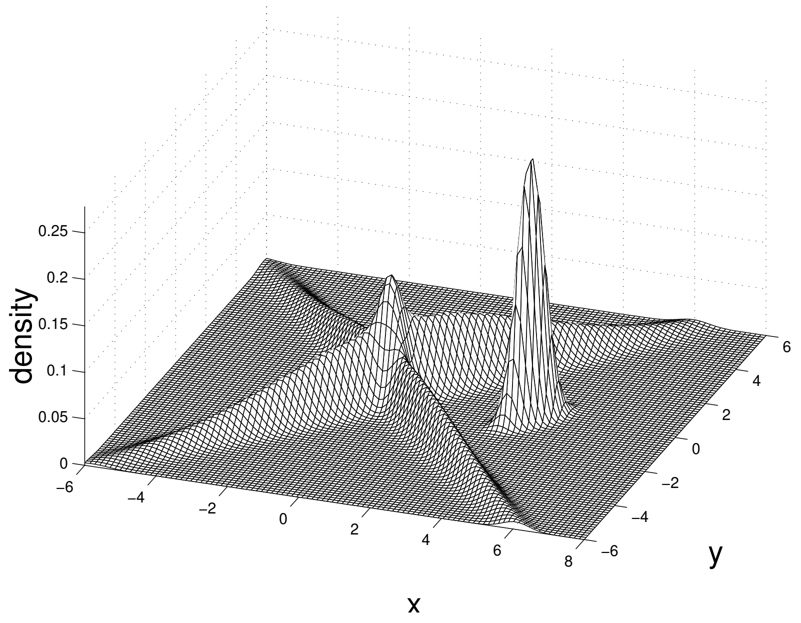
$$CL(\theta, z) = \sum_{k=1}^{K} \sum_{i/z_i=k} \ln(p_k \phi(x_i | a_k))$$

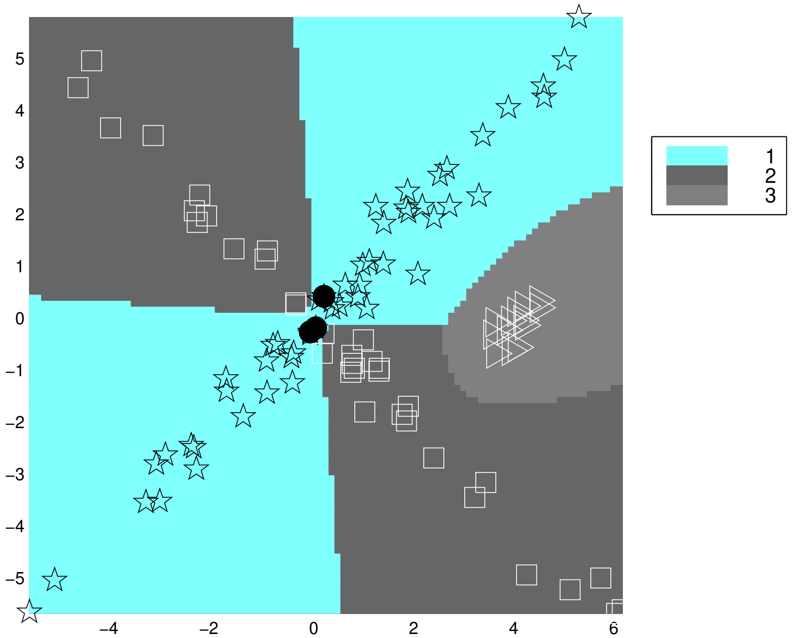
2. $t_k(x_i)$ : probabilité que $x_i$ provienne du groupe $k$
3. $\hat{z}$ : chaque $x_i$ est affecté à la classe la plus probable (MAP)

# Algorithmes

- estimateur MV : algorithmes EM, SEM
- estimateur MVC : algorithme CEM (avantages pratiques)

## Exemple avec EM

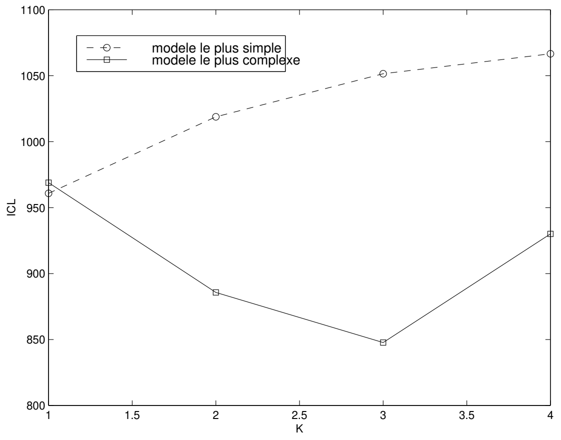


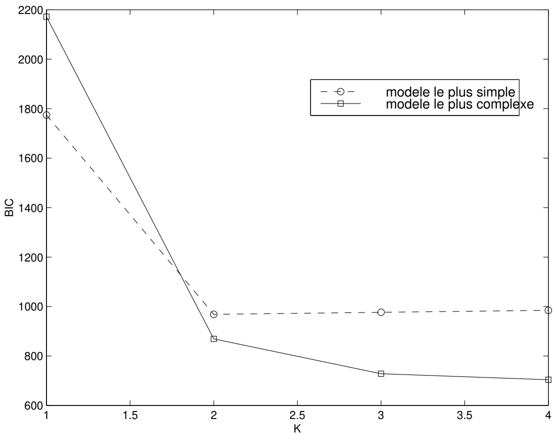
# Choix d'un modèle et du nombre de classes

- ***utilisateur*** : connaissances *a priori*
- ***critères BIC, AIC*** : pénalisation de la vraisemblance $L - f(taille\ \theta)$
- ***critères ICL, NEC*** : favorisent les classes séparées

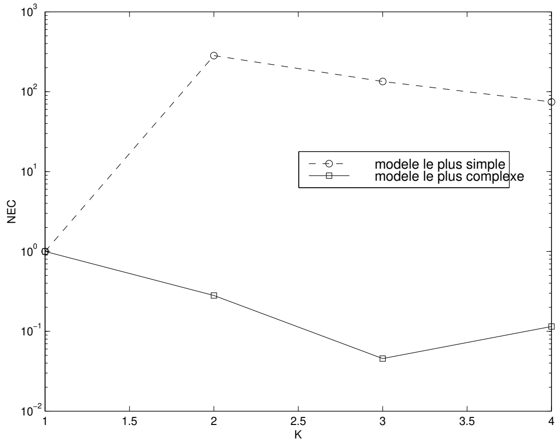
# Exemple




ICL




BIC




NEC

# LA CLASSIFICATION AVEC CONNAISSANCE PARTIELLE

## Les données

$$x = (x', x'') \qquad z = (z', \boxed{z''}) \text{ avec } z'' \text{ inconnue}$$

## Estimation des labels $z''$

Les critères L et CL ainsi que les algorithmes EM, SEM, CEM peuvent être adaptés facilement à cette situation.

## Critères de choix de modèles et de $K$

Ils sont encore utilisables.

# LA DISCRIMINATION GÉNÉRALISÉE

## Le problème

***Petrel adulte*** variable biometrique x

sexe z

***Jeune petrel*** variable biometrique x*

sexe z*

inconnu

## Les données

- $n$ oiseaux adultes (population $P$) : variables biométriques $x$ et sexes $z$
- $n^*$ jeunes oiseaux (population $P^*$) : (mêmes) variables biométriques $x^*$ et sexes $z^*$ ***inconnus***

## Objectif

$$\left.\begin{array}{ll} \text{données de } P : & x,\, z \\ \text{données de } P^* : & x^* \end{array}\right\} \longrightarrow \hat{z}^* \text{ de } P^*$$

## Principe

1. établir une relation linéaire entre les variables de même sexe mais de population différente

$$X^*_{|Z^*=k} \sim D_k X_{|Z=k} \qquad (D_k \text{ diagonale})$$

2. estimer $D_k$ et déduire une estimation des paramètres de $P^*$

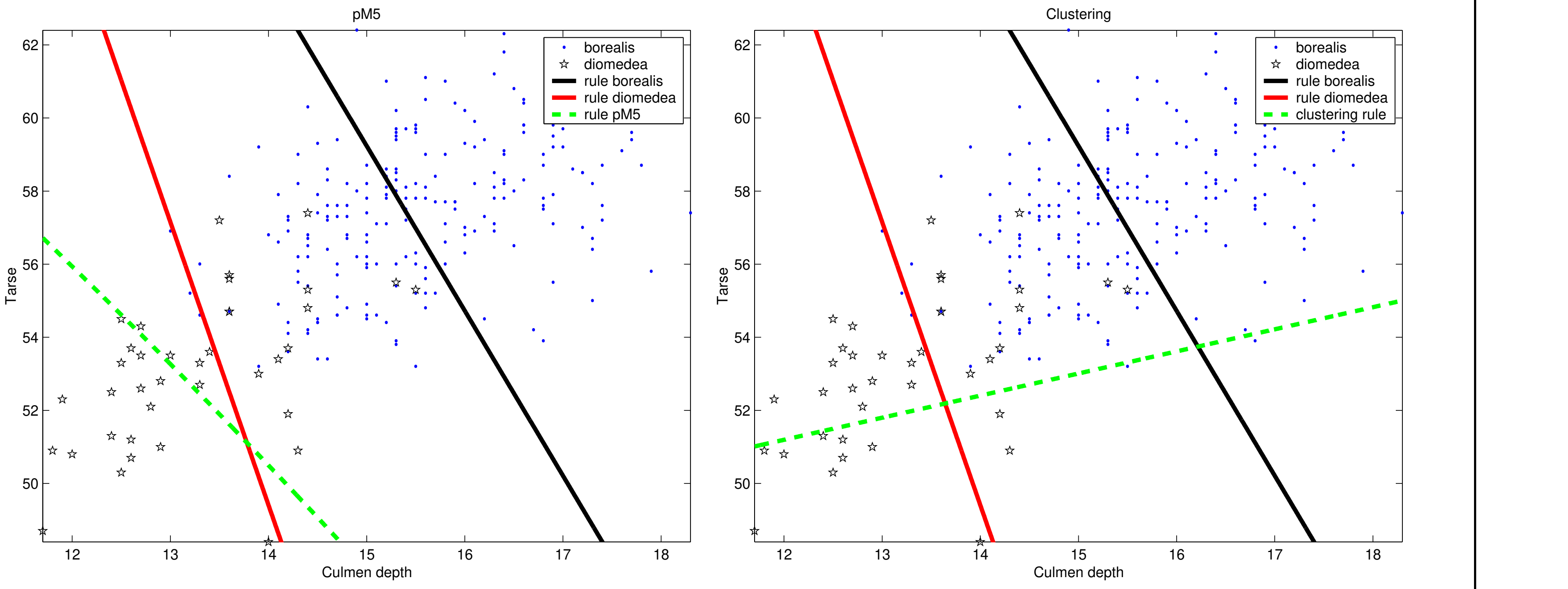
3. une estimation de la règle de classement sur $P^*$ est immédiate

## Remarque

Il y a peu de paramètres à estimer.

# Exemple avec une différence géographique

- $P$ : petrels atlantiques des Açores (espèce borealis)
- $P^*$ : petrels méditérannéens des Baléares (espèce diomedea)



# CONCLUSION

- classification et discrimination : problèmes concrets
- les mélanges : cadre probabiliste rigoureux et performant
- implantation des méthodes : relativement simple
- utilisation : facilité par des interprétations géométriques