# Panorama des logiciels d'identification des mélanges

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

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

# Plan

**MCLUST**

**EMMIX**

**AUTOCLASS**

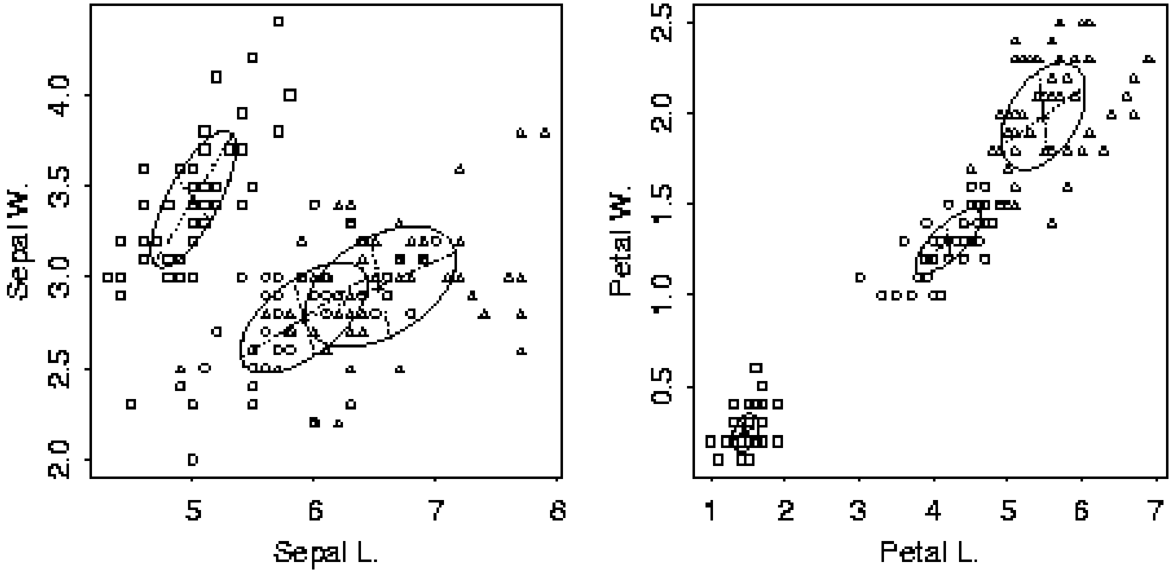
**C.A.MAN**

**XEMgaus**

**Bilan**

**Conclusion**

# MCLUST



Sepal W.
2.0
2.5
3.0
3.5
4.0
5
6
7
8
Sepal L.
Petal W.
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Petal L.

# Identification

- auteurs : C. Fraley et A.E. Raftery (university of Washington, USA)
- année : 1998
- `http://www.stat.washington.edu/fraley/software.html`
- langage : fortran interfacé avec le logiciel Splus (ou R)
- systèmes : unix, windows
- diffusion : gratuit
- documentation : théorie + exemples

# Caractéristiques liées aux mélanges

- objectif : classification et discrimination
- algorithmes :
  - EM pour modèles gaussiens multivariés avec possibilité d'ajouter un bruit poissonnien
  - classification hiérarchique avec modèles gaussiens multivariés
- initialisation de EM : par classification hiérarchique
- modèles disponibles (6/14) : $[\lambda I]$, $[\lambda_k I]$, $[\lambda C]$, $[\lambda_k C_k]$, $[\lambda D_k A D_k]$, $[\lambda_k D_k A D_k]$
- critères : BIC
- possibilité de lancer plusieurs modèles et plusieurs nombres de classes en une seule fois
- graphiques : partitions et ellipses de variance sous Splus (ou R)

## EMMIX

**EM based MIXture program**

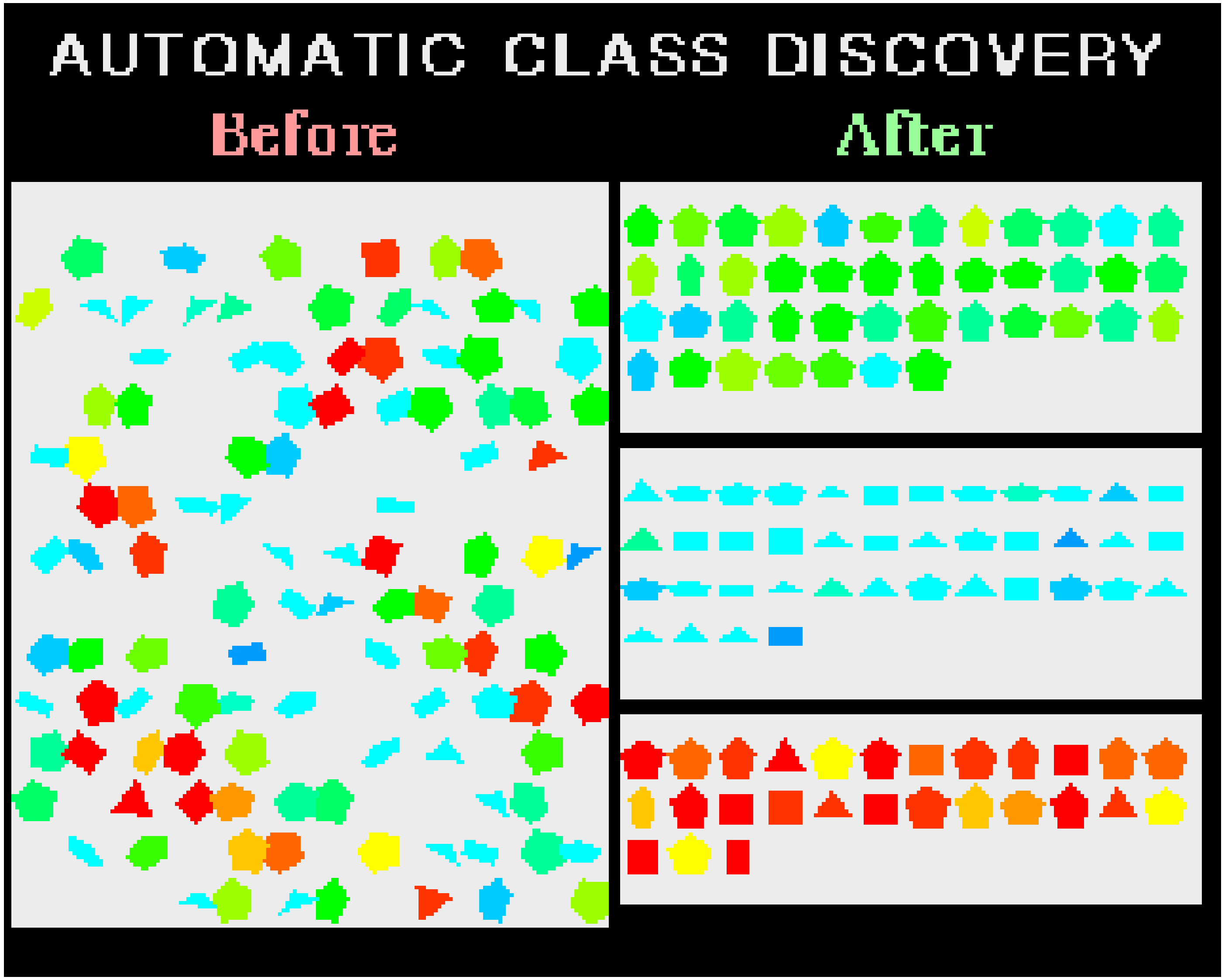
## Identification

- auteurs : D. Peel et G.J. McLAchlan (university of Queensland, Australia)
- année : 1998
- `http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`
- langage : fortran
- systèmes : unix, windows
- diffusion : gratuit
- documentation : théorie et exemples

## Caractéristiques liées aux mélanges

- objectif : classification, discrimination, classification partielle
- algorithmes : EM et SEM pour modèles gaussiens multivariés et pour modèles de Student multivariés
- initialisation de EM : utilisateur (partition, paramètre), automatique (classification hiérarchique, $k$means, partition aléatoire)
- modèles disponibles (4/14) : $[\lambda I]$, $[\lambda_k C_k]$, $[\lambda_k B_k]$, $[\lambda B]$
- critères : BIC, AIC, AWE, P-value, bootstrap paramétrique
- arrêt EM : nombre d'itérations, quasi-stabilité de la vraisemblance sur les 10 dernières itérations
- possibilité de lancer plusieurs nombres de classes en une seule fois
- exportation des résultats vers des logiciels de visualisation
- estimation de l'écart-type des estimateurs par bootstrap

# AUTOCLASS



# Identification

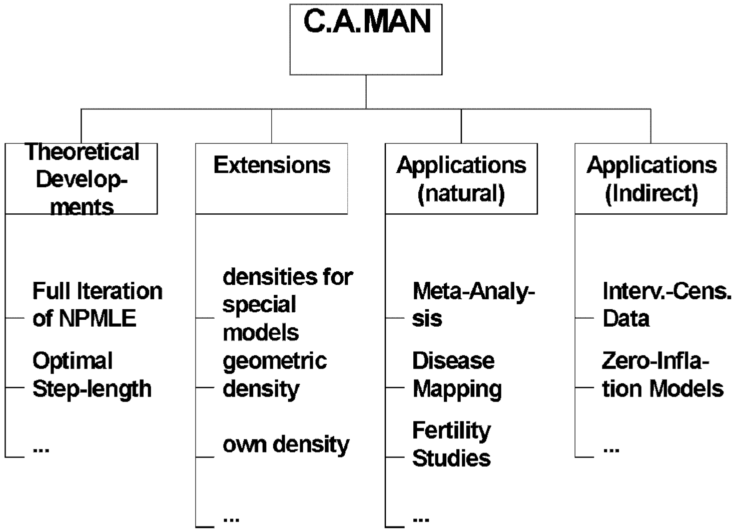
- auteurs : P. Cheeseman (RIACS, USA) et J. Stutz (NASA, USA)
- année : 1995 (1ère version), 2000 (dernière version)
- `http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/`
- langage : C
- systèmes : unix, windows
- diffusion : gratuit
- documentation : théorie

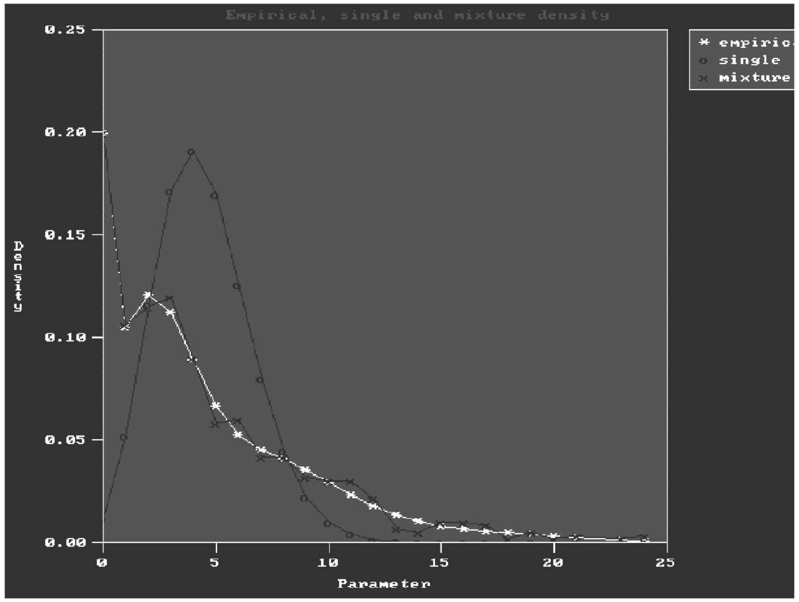
# Caractéristiques liées aux mélanges

- objectif : classification et discrimination bayésiennes
- algorithmes : bayésiens avec différents modèles
  - mélange gaussien pour données quantitatives
  - mélange de Bernoulli pour données qualitatives
  - mélange de Poisson pour données de comptage
  - possibilité de mixer gaussien et Bernoulli
- modèles gaussiens disponibles : $[\lambda_k B_k]$, $[\lambda_k C_k]$ (2/14) + covariances par blocs
- modèles de Bernoulli disponibles : avec ou sans corrélation
- critère bayésien de choix du nombre de classes
- gestion des valeurs manquantes
- difficulté : choix des paramètres de lois a priori

# C.A.MAN

## Computer Assisted Analysis of Mixtures



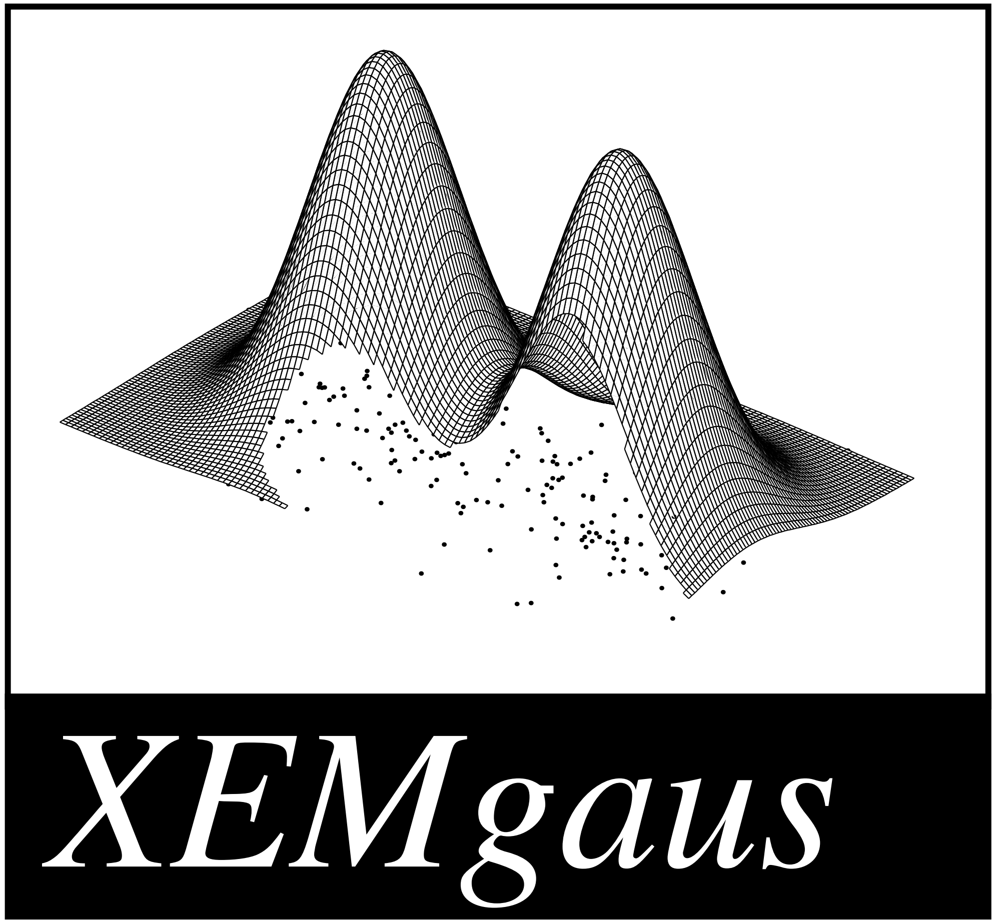
## Identification

- auteurs : D.B. Böhning et P. Schlattmann (université libre de Berlin, Allemagne)
- année : 1997
- `http://www.medizin.fu-berlin.de/sozmed/caman.html`
- langage : Fortran
- systèmes : unix, windows
- diffusion : gratuit
- documentation : quasi inexistante

## Caractéristiques liées aux mélanges

- objectif : modèles de mélanges
- algorithmes : VEM, VEM-EM, VDM, EM pour le maximum de vraisemblance non paramétrique
- modèles disponibles :
  - gaussiens (2/14, unidimensionnels) : $[\lambda I]$, $[\lambda_k I]$
  - autres : Poisson, Laplace, exponentiel, binomial, géométrique
- fonctions graphiques
- limitation : 500 points

# XEMGAUS



## Identification

- auteurs :
  - G. Celeux et Y. Vernaz (Inria Rhône-Alpes)
  - C. Biernacki et F. Langrognet (université de Franche-Comté)
  - C. Ambroise et G. Govaert (université de technologie de Compiègne)
- année : juin 2001
- version Matlab actuelle : `http://www.inrialpes.fr/is2/`
- langage : C++ interfacé avec Scilab
- systèmes : unix, windows
- diffusion : gratuit
- documentation : théorie + exemples

# Caractéristiques liées aux mélanges

- objectif : classification, discrimination, classification partielle
- algorithmes : EM, CEM, SEM pour les mélange gaussiens
- modèles gaussiens disponibles : les 14
- possibilité d'enchaîner les modèles et les algorithmes
- critères : BIC, ICL, NEC, validation croisée
- fonctions graphiques : interfaçage prévu avec Scilab

# BILAN

## Algorithmes

| | MCLUST | EMMIX | AUTOCLASS | C.A.MAN | XEMgaus |
|---|---|---|---|---|---|
| EM | × | × | | × | × |
| SEM | | × | | | × |
| CEM | | | | | × |
| hiérarchique | × | | | | |
| Bayes | | | × | | |
| autres | | | | × | |
| Enchaînements | | | | × | × |

# Distributions

| | MCLUST | EMMIX | AUTOCLASS | C.A.MAN | XEMgaus |
|---|---|---|---|---|---|
| gauss. univarié | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ |
| gauss. multivarié | $\times$ | $\times$ | $\times$ | | $\times$ |
| binomial | | | $\times$ | $\times$ | |
| Poisson | | | $\times$ | $\times$ | |
| exponentiel | | | | $\times$ | |
| géométrique | | | | $\times$ | |
| Laplace | | | | $\times$ | |
| Student | | $\times$ | | | |
| bruit | $\times$ | | | | |
| combinaisons | | | $\times$ | | |

# Critères

| | MCLUST | EMMIX | AUTOCLASS | C.A.MAN | XEMgaus |
|---|---|---|---|---|---|
| BIC | $\times$ | $\times$ | | | $\times$ |
| AIC | | $\times$ | | | |
| ICL | | | | | $\times$ |
| NEC | | | | | $\times$ |
| AWE | | $\times$ | | | |
| val. crois. | | | | | $\times$ |
| P-value | | $\times$ | | | |
| bootstrap | | $\times$ | | | |
| Bayes $\neq$ BIC | | | $\times$ | | |

## Objectifs

| | MCLUST | EMMIX | AUTOCLASS | C.A.MAN | XEMgaus |
|---|---|---|---|---|---|
| mélange seul. | × | × | × | × | × |
| classif. auto. | × | × | × | | × |
| discrimination | × | × | × | | × |
| classif. partielle | | × | | | × |

## Modèles gaussiens

| MCLUST | EMMIX | AUTOCLASS | C.A.MAN | XEMgaus |
|---|---|---|---|---|
| 6 | 4 | 2 | 2 | 14 |

# Conclusion

## Choix d'un logiciel

- en fonction de l'objectif et du modèle
- ouvert sur des outils évolués comme Scilab, R : XEMgaus, MCLUST

## Autres logiciels

Une liste de logiciels est disponible à l'adresse

`http://www.csse.monash.edu.au/~dld/cluster.html`