# Intégration de données qualitatives et quantitatives par les modèles de mélange

**Christophe Ambroise**

UMR CNRS 6599, Université de Technologie de Compiègne

**9 mars 2001**

# Plan

1. Introduction
2. Quelques solutions
3. Application
4. Conclusion

# Introduction

- Modèles de mélange fini pour :
  - Discrimination
  - Clustering
- Type de données :
  - Variables quantitatives
  - Variables qualitatives
- Comment intégrer le mélange de données qualitatives et quantitatives ?

## Types de données

- Variables quantitatives

  Les variables quantitatives prennent leur valeurs sur $\mathcal{R}$ :

  – continue

  – discrète

| | Variables | |
|---|---|---|
| | taille | poids |
| Pierre | 1.80 | 73 |
| Pauline | 1.60 | 55 |
| Jacques | 1.75 | 65 |

- Variables qualitatives

  Les variables qualitatives prennent leur valeurs sur un ensemble fini de modalités :

  – nominale

  – ordinale

| | Variables | |
|---|---|---|
| | Sexe | Revenu |
| Pierre | H | niveau1 |
| Pauline | F | niveau3 |
| Jacques | H | niveau10 |

## Notations

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

- Ensemble des individus: $(\mathbf{w}_1,\ldots,\mathbf{w}_n)$ avec $\mathbf{w}_i=(\mathbf{x}_i,\mathbf{y}_i)$, et
  - $\mathbf{x}_i=(\mathbf{x}_i^1,\ldots,\mathbf{x}_i^c)$ sont les $c$ variables qualitatives,
  - $\mathbf{y}_i=(y_i^1,\ldots,y_i^d)$ sont les $d$ variables quantitatives.
- $\mathbf{x}_i^j=(x_i^{j1},\ldots,x_i^{jm_j})$, codage binaire des $m_j$ modalités
  Exemple $\mathbf{x}_i^j=(001)$
- Nombre de classes: $k=1,\ldots,g$,
- $f_k(\mathbf{w}_i;\theta_k)$, densité de la classe $k$,
- $z_i$, vecteur binaire indicateur de la classe de $\mathbf{x}_i$

# Quelques solutions

- Transformation de variables
- Prise en compte des différences de nature
  - Indépendance,
  - Modèle complet,
  - Dépendance conditionnelle,

## Transformation de variables

- Variables qualitatives $\rightarrow$ Variables quantitatives
  - Attention :transformation des variables qualitatives nominales, en variables qualitatives ordinales.
  - Codage binaire, modèle de Bernouilli.
- Variables quantitatives $\rightarrow$ Variables qualitatives
  - Inconvénient : Explosion du nombre de paramètres.

## Modèle supposant l'indépendance

- Hypothèses :
  - Indépendance entre quantitatif et qualitatif,

$$f_k(\mathbf{w}_i;\theta_k) = f_k(\mathbf{x}_i,\mathbf{y}_i;\theta_k) = f_{kx}(\mathbf{x}_i;\theta_{kx})f_{ky}(\mathbf{y}_i;\theta_{ky})$$

  - Indépendance entre les variables qualitatives (Classes Latentes),

$$f_k(\mathbf{w}_i;\theta_k) = \left(\prod_{j=1}^{c} f_{kx}(\mathbf{x}_i^j;\theta_{kj})\right) f_{ky}(\mathbf{y}_i;\mu_k,\Sigma_k)$$

  - $f_{kx}(\mathbf{x}_i^j;\theta_{kj}) = \prod_{s=1}^{m_j} (\pi_k^{js})^{x_i^{js}}$ est une loi multinomiale avec $m_j$ modalités.
  - $f_{ky}(\mathbf{y}_i;\mu_k,\Sigma_k)$ est une loi normale
- Remarque : modèle complet exploitable grâce aux variables quantitatives.

## Modèle supposant l'indépendance

- Implémentation du modèle :

**E-step**

$$Q(\Phi|\Phi^q) = \sum_{ik} t_{ik}^q \mathrm{E}_Z \left[\log p_k f_{kx}(\mathbf{x}_i;\theta_{kx}) f_{ky}(\mathbf{y}_i;\theta_{ky})|\mathbf{w}_i\right],$$

avec

$$t_{ik}^q = \frac{p_k^q \prod_{j=1}^c \prod_{s=1}^{m_j} \left(\pi_k^{js}\right)^{x_i^{js}} f_{ky}(\mathbf{y}_i;\mu_k^q,\Sigma_k^q)}{f(\mathbf{w}_i;\Phi^q)}$$

**M-step** Maximisation de $Q(\Phi|\Phi^q)$

$$\pi_k^{js} = \frac{\sum_i t_{ik}^q x_i^{js}}{\sum_i \sum_{s=1}^{m_j} t_{ik}^q x_i^{js}} \qquad \mu_k = \frac{\sum_i t_{ik}^q \mathbf{y}_i}{\sum_i t_{ik}^q}$$

## Modèle supposant l'indépendance

- Avantages : nombre de paramètres raisonnable

  –

| Nb de paramètres | Exemple (3 classes) |
|---|---|
| $\sum_{k=1}^{g} \sum_{j=1}^{c} m_j$ proportions qualitatives | 36 |
| $g$ vecteurs moyennes | 6 |
| $g$ matrices de covariances | 9 |
| | **51** |

- Inconvénients : vraisemblance des hypothèses
  - indépendance quantitatif/qualitatif :
    la taille est-elle indépendante du sexe?
  - indépendance entre variables qualitatives :
    le niveau de revenu est il indépendant du sexe?

## Dépendances entre variables qualitatives

- Hypothèses :
  - Indépendance entre quantitatif et qualitatif,

$$f_k(\mathbf{w}_i;\theta_k) = f_k(\mathbf{x}_i,\mathbf{y}_i;\theta_k) = f_{kx}(\mathbf{x}_i;\theta_{kx})f_{ky}(\mathbf{y}_i;\theta_{ky})$$

  - Dépendance entre variable qualitatives (modèle complet) :

*création d'une variable unique à* $m = \prod_{j=1}^{c} m_j$ *modalités.*

| | niveau1 | niveau2 | ... | niveau9 | niveau10 |
|---|---|---|---|---|---|
| Homme | | | ... | | |
| Femme | | | ... | | |

## Dépendances entre variables qualitatives

- Avantages : nombre de paramètres raisonnable
  –

| Nb de paramètres | Exemple (3 classes) |
|---|---|
| $\sum_{k=1}^{g} \prod_{j=1}^{c} m_j$ proportions qualitatives | 60 |
| $g$ vecteurs moyennes | 6 |
| $g$ matrices de covariances | 9 |
| | **75** |

- Inconvénients : vraisemblance des hypothèses
  - indépendance quantitatif/qualitatif :
    la taille est-elle indépendante du sexe?
  - Toutes les modalités de la variable complète sont-elles présentes?

## Dépendances conditionnelles

- Hypothèses :
  - Dépendance entre variable qualitatives :
    création d'une variable unique à $m = \prod_{j=1}^{c} m_j$ modalités.
  - Dépendance conditionnelle entre quantitatif et qualitatif,

$$f_k(\mathbf{w}_i;\theta_k)$$

$$\swarrow \qquad\qquad\qquad \searrow$$

$$f_{kx}(\mathbf{x}_i) f_{ky}(\mathbf{y}_i|\mathbf{x}_i) \qquad f_{ky}(\mathbf{y}_i) f_{kx}(\mathbf{x}_i|\mathbf{y}_i)$$

  -

## Dépendances conditionnelles

Location Based Model

-
$$
\begin{aligned}
f_k(\mathbf{w}_i;\theta_k) &= f_{kx}(\mathbf{x}_i;\theta_{kx})f_{ky}(\mathbf{y}_i|\mathbf{x}_i;\theta_{ky}) \\
&= \prod_{s=1}^{m} \left(\pi_k^s \cdot f_{ky}(\mathbf{y}_i|x_i^s = 1;\mu_{ks},\Sigma_{ks})\right)^{x_i^s}
\end{aligned}
$$

avec $f_{ky}(\mathbf{y}_i|x_i^s = 1;\mu_{ks},\Sigma_{ks})$ une loi normale

## Dépendances conditionnelles

- Implémentation du modèle :

**E-step** Calcul de $Q(\Phi|\Phi^q)$ avec

$$t^q_{iks} = \frac{p^q_k \pi^s_k f_{ky}(\mathbf{y}_i | x^s_i = 1; \mu_{ks}, \Sigma_{ks})}{\sum_{\ell=1}^{g} p^q_\ell \pi^s_\ell f_{\ell y}(\mathbf{y}_i | x^s_i = 1; \mu_{\ell s}, \Sigma_{\ell s})}$$

**M-step** Maximisation de $Q(\Phi|\Phi^q)$

$$\pi^s_k = \frac{\sum_i t^q_{iks} x^s_i}{\sum_i \sum_{s=1}^{m} t^q_{iks} x^s_i} \qquad \mu^{q+1}_{ks} = \frac{\sum_i t^q_{iks} \mathbf{y}_i}{\sum_i t^q_{iks}}$$

## Dépendances conditionnelles

- Avantages :
  - prise en compte d'hypothèses complexes
- Inconvénients : nombre de paramètres
  -

| Nb de paramètres | Exemple (3 classes) |
|---|---|
| $\sum_{k=1}^{g} \prod_{j=1}^{c} m_j$ proportions qualitatives | 60 |
| $g$ vecteurs moyennes | 120 |
| $g$ matrices de covariances | 180 |
| | **360** |

# Application

- Benchmark proposé par COPARLY
- Question : Discriminer un jour J pollué à partir d'informations disponibles le jour J-1 et d'éventuelles prévisions?
- Les données sur Lyon
  - 5 Années de 1994 à 1998
  - Périodes du 1 Avril au 30 Septembre 915 jours de données

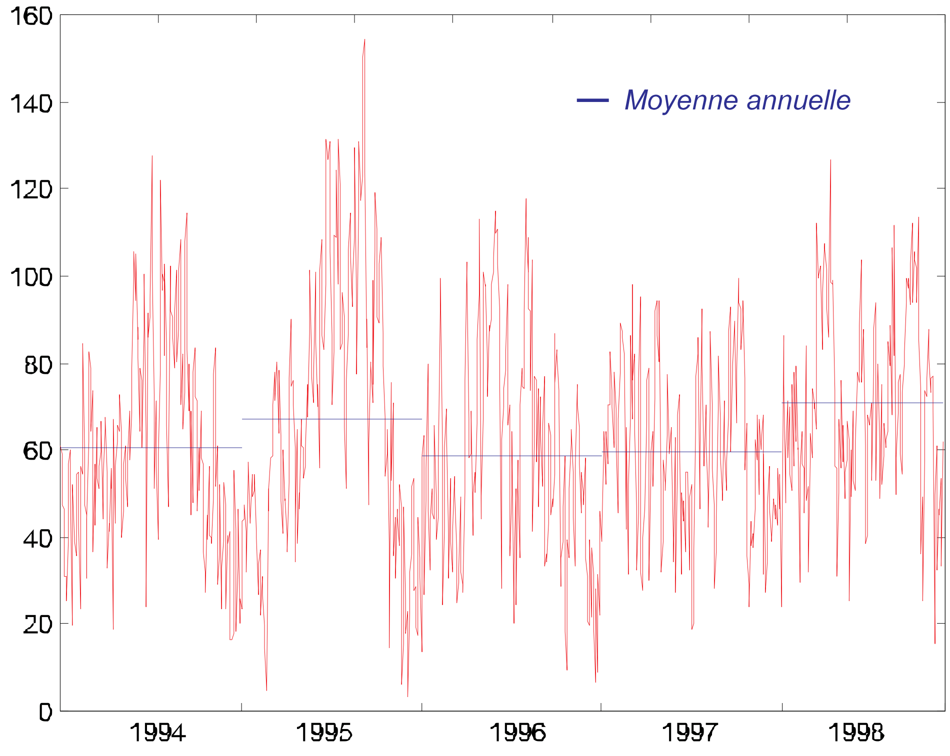
# Le site d'étude

# Les informations disponibles

- Pollution
  - 16 + 1 capteurs $NO$, $NO_2$ (11% données manquantes)
  - 7 capteurs $O_3$ (24% données manquantes)
- Météo
  - Température : Mesure horaire et Prévision à 24H
  - Géopotentiel de Bénichou : Prévision journalière des conditions atmosphériques (1.4% données manquantes)

# Analyse préliminaire des données

➠ Moyenne journalière $O_3$



# Particularités des données

- Nombreux attributs corrélés (1012 pour 900 exemples !)
- Evènements de pollution rares
- Variables qualitatives et quantitatives
- Données non-validées
- Attributs et étiquettes manquants
  - Evolution du réseau de mesures
  - Capteurs défectueux

# Discrimination par modèle de mélange

- Variables qualitatives :
  - Classe de Benichou 36 heures avant la prévision (3 modalités)
  - Mois de l'année (6 modalités: avril à septembre)
- Variables quantitatives :
  - Température prévue
  - Force du vent (24 heure avant la prévision)
  - Concentration maximum d'ozone (le jour avant la prévision)

# Discrimination par modèle de mélange

- 3 tentatives de classification :

1. variables qualitatives seules

| | 1994-1997 (apprentissage) PRÉVISION | | 1998 (test) PRÉVISION | |
|---|---|---|---|---|
| OBSERVATION | pollution | normal | pollution | normal |
| pollution | 5.87 % | 0.41 % | 6.01 % | 0.55 % |
| normal | 42.21 % | 51.50 % | 41.53 % | 51.91% |

2. variables quantitatives seules

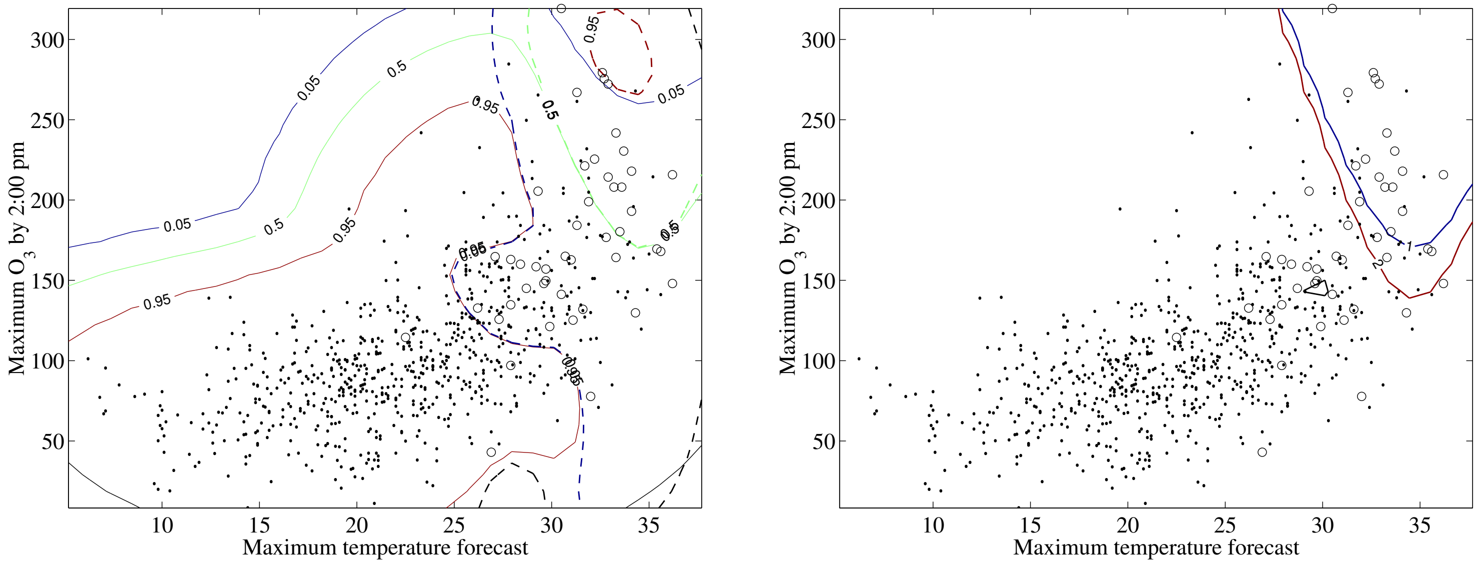
| | 1994-1997 (apprentissage) PRÉVISION | | 1998 (test) PRÉVISION | |
|---|---|---|---|---|
| OBSERVATION | pollution | normal | pollution | normal |
| pollution | 4.10 % | 2.19 % | 4.37 % | 2.19% |
| normal | 10.38 % | 83.33 % | 8.74% | 84.70% |

3. toutes les variables (hypothèses d'indépendances)

| | 1994-1997 (apprentissage) PRÉVISION | | 1998 (test) PRÉVISION | |
|---|---|---|---|---|
| OBSERVATION | pollution | normal | pollution | normal |
| pollution | 2.73 % | 3.42 % | 1.0 % | 5.46 % |
| normal | 3.42 % | 90.30 % | 2.19 % | 91.25 % |

# Frontières de décision, modèle de mélange

Marginalisation sur les deux variables principales :



# Remarques

- Intérêt de la prise en compte des natures différentes des variables;
- Problème du nombre de paramètres:
  - Modèle complexe peu réaliste
  - Complexité via le nombre de classes/Complexité via les hypothèses de dépendance.
- Importance des variables:
  - quantitatives/quantitatives
  - qualitatives/qualitatives
  - qualitatives/quantitatives