Modèles de mélange sur variable qualitatives et binaires

Gérard Govaert et Gilles Celeux

9 mars 2001

I Introduction

Objectif : prise en compte des variables qualitatives dans les modèles de mélange

- Variables binaires : variables qualitatives à 2 modalités
- Variables qualitatives
- Variables qualitatives ordinales
- Variables quantitatives et variables qualitatives

Variables binaires

- Courantes en statistique
- Données de présence-absence (écologie, par exemple)
- Image : pixels noirs ou blancs
- Recodage de variables qualitatives
- Données :
  - échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$
  - $\mathbf{x}_i = (x_i^1, \ldots, x_i^d)$
  - $x_i^j \in \{0, 1\}$ (éventuellement après recodage)

- Différents codages :
  - tableau individus variables : n, d
  - fréquences des états
  - table de contingence
- Exemple : 10 parcelles sur lesquelles on a observé (1) ou non (0) la présence de 2 papillons P1, P2 :

Parcelle	P1	P2
1	1	1
2	1	0
3	1	1
4	0	1
5	0	1
6	0	0
7	1	0
8	1	1
9	1	0
10	0	0

Etat	Fréquence
00	2
01	2
10	3
11	3

P1 \ P2	0	1
0	2	2
1	3	3

- Remarques :
  - En analyse de données, ce type de données a donné lieu à un grand nombre de distances :

$i \backslash i'$	1	0
1	a	b
0	c	d

  - $d(i, i') = a + d$: coefficient d'accord (matching coefficient)
  - $d(i, i') = a/(a + b + c)$ (indice de Jaccard)
  - Symétrie des 2 valeurs binaires ou non

Modèle de mélange binaire

- Distribution mélange : $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ réalisation d'un vecteur aléatoire de $\{0,1\}^d$ de densité

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

La classe k est caractérisée par la distribution

$$f_k(\mathbf{x};\theta_k)$$

- Choix de f_k :
  - Modèles classiques :
    - Quantitatif : modèle gaussien
    - Qualitatif : modèle log-linéaire
  - Modèle complet ou saturé : chaque classe suit une distribution multinomiale à 2^d valeurs : aucun intérêt dans le cadre d'un mélange
  - Il faut prendre un modèle log-linéaire suffisamment contraint :

Modèle d'indépendance

- Modèle des classes latentes : Lazarfield and Henry (1968), Goodman (1974)
  - Indépendance conditionnelle
  - x^j suit une distribution de Bernoulli :

$$f(x^j;\theta_k^j) = (\theta_k^j)^{x^j}(1-\theta_k^j)^{1-x^j} \text{ où } \theta_k^j = Pr(x^j = 1/k)$$

  - Distribution de la classe k :

$$f_k(\mathbf{x};\theta_k) = \prod_{j=1}^{d}(\theta_k^j)^{x^j}(1-\theta_k^j)^{1-x^j} \text{ où } \theta_k = (\theta_k^1,\ldots,\theta_k^d)$$

- Distribution mélange :
  - $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ réalisation d'un vecteur aléatoire de $\{0,1\}^d$ de densité

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

  - Paramètre $\Phi = (\mathbf{p}, \theta)$:
    * Proportions $\mathbf{p} = (p_1, \ldots, p_g)$
    * Paramètres de chaque composant $\theta = (\theta_1, \ldots, \theta_g)$

Problème :

- Estimer le paramètre

$$\Phi = (\mathbf{p}, \theta) = (p_1, \ldots, p_g, \theta_1^1, \ldots, \theta_g^d)$$

- Eventuellement, estimer la classe d'origine
- Solution : maximiser la vraisemblance
- Nombre de paramètres :
  - Classes latentes à g classes: $(g-1) + g * d$
  - Modèle complet : 2^d
  - pour g=5 et d=10 : 54 et 1024
- Conditions d'identifiabilité :

$$\text{nombre d'états} \; > \; \text{nombre de paramètres}$$

Vraisemblances

- Vraisemblance

$$\begin{aligned} L(\Phi) &= L(\Phi;\mathbf{x}) = \sum_{i=1}^{n}\log\{\sum_{k=1}^{g} p_k f_k(\mathbf{x}_i;\theta_k)\} \\ &= \sum_{i=1}^{n}\log\{\sum_{k=1}^{g} p_k \prod_{j=1}^{d}(\theta_k^j)^{x_i^j}(1-\theta_k^j)^{1-x_i^j}\} \end{aligned}$$

- Vraisemblance complétée

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

Algorithme EM

- Étape 0 : solution initiale $\Phi^{(0)} = (\mathbf{p}^{(0)}, \theta^{(0)})$
- 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 M :
  - $p_k^{(c+1)} = \frac{\sum_i t_{ik}}{n}$
  - $(\theta_k^j)^{(c+1)} = \frac{\sum_{i=1}^{n} t_{ik} x_i^j}{\sum_{i=1}^{n} t_{ik}}$

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 : $\mathbf{z}^{(c+1)} = \text{MAP}(t_{ik})$:
  Les t_{ki} sont remplacées par des 1 ou des 0
- Étape M : maximisation de $L_C(\mathbf{z}^{(c+1)}, .)$

  – $p_k^{(c+1)} = \frac{\# z_k}{n}$
  – $(\theta_k^j)^{(c+1)} = \frac{\sum_{i=1}^{n} t_{ik} x_i^j}{\# z_k} =$ % de 1

Modèle binaire parcimonieux : reparamétrisation

$$f(\mathbf{x};\Phi) = \sum_{k=1}^{g} p_k \prod_{j=1}^{d} (\varepsilon_k^j)^{|x^j - a_k^j|}(1-\varepsilon_k^j)^{1-|x^j-a_k^j|}$$

$$\begin{cases} a_k^j = 0, \varepsilon_k^j = \theta_k^j & \text{si } \theta_k^j < 0.5 \\ a_k^j = 1, \varepsilon_k^j = 1-\theta_k^j & \text{si } \theta_k^j > 0.5 \end{cases}$$

$$\begin{cases} \text{centre } \mathbf{a}_k \in \{0,1\}^d & = & \text{valeurs } 0/1 \text{ les plus probables} \\ \text{dispersion } \varepsilon_k \in]0,1/2[^d & = & \text{probabilité d'être différent du centre} \end{cases}$$

- Paramétrage utilisé pour la discrimination avec estimation non paramétrique par la méthode des noyaux (Aitchinson and Aitken, 1976)
- Modèle initial : ε_k^j
- Modèle parcimonieux : ε, ε_k, ε^j
- Calcul du nombre de paramètres ?

Expression de la vraisemblance complétée

On se limite au cas le plus simple : modèle ε et $p_k = 1/g$

on a alors

$$L_C(\mathbf{z}, \Phi) = \log(\frac{\varepsilon}{1-\varepsilon}) \sum_{k=1}^{g} \sum_{i \in z_k} d(\mathbf{x}_i, \mathbf{a}_k) + nd \log(1-\varepsilon)$$

où

$$d(\mathbf{x}_i, \mathbf{a}_k) = \sum_{j=1}^{d} |x_i^j - a_k^j|$$

On retrouve ainsi une métrique souvent utilisée pour des données binaires (coefficient d'accord)

Algorithme CEM correspondant

- Étape 0 : solution initiale $(\Phi^0, \mathbf{z}^0)$
- Étape E : les individus i sont rangés dans la classe k qui minimise

$$d(\mathbf{x}_i, \mathbf{a}_k)$$

- Étape M :
  - $p_k^{(c+1)} = \frac{\# z_k}{n}$
  - $(a_k^j)^{(c+1)} =$ valeur binaire majoritaire de la classe k pour j

Exemple d'application

Données de Stouffer et Tobby (1951), $n = 216$, $d = 4$

$S1$	$S2$	$S3$	$S4$	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

- Résultats :
  - Résultats EM sur le modèle des classes latentes (9 paramètres) :

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

    Déviance : 2.72
  - Modèle log-linéaire avec interaction d'ordre 2 (11 paramètres) : deviance de 7.11

Discussion

- Choix du modèle et du nombre de classes
- Approche mélange-approche classification

Variables qualitatives

- Questionnaire, données discrètes, données catégorielles
- Généralise les données binaires à plus de 2 modalités
- Données :
  - Échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$
  - $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

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

  - m_j nombre de modalités de la variable qualitative j
- Mêmes codages que pour les binaire : tableau individus variables, fréquences des états ou table de contingence

Un exemple : les félins

No	Nom	Libelle
1	type	aspect du pelage
	typ1	sans tache
	typ2	tachete
	typ3	raye
	typ4	marbre
2	poil	fourrure
	poi1	poils ras
	poi2	poils longs
3	grif	griffes retractiles
	gri1	retractiles
	gri2	non retractiles

	type	*poil*	*grif*
1	1	1	2
2	3	1	2
3	2	1	2
4	2	1	2
5	2	2	2
6	2	1	1
7	1	1	2
8	4	1	2
9	2	1	2
10	2	1	2
11	2	2	2
12	1	1	2
13	2	1	2
14	1	1	2
15	1	2	2
16	1	1	2
17	2	1	2
18	1	2	2
19	3	1	2
20	1	1	2

Modèle de mélange qualitatif

- Comme pour le binaire : modèle des classes latentes
  - Indépendance conditionnelle
  - x^j suit une distribution multinomiale

$$f(x^j;\theta_k^j) = \prod_{h=1}^{m_j} (\theta_k^{jh})^{x^{jh}}$$

où $\theta_k^{jh} = Pr(x^j = h/k)$ avec $\sum_h \theta_k jh = 1$

  - Distribution f_k :

$$f(\mathbf{x};\theta_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\theta_k^{jh})^{x^{jh}}$$

où $\theta_k = (\theta_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

- Distribution mélange :
  - $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ réalisation d'un vecteur aléatoire de densité

$$f(\mathbf{x}; \Phi) = \sum_{k=1}^{g} p_k f(\mathbf{x}; \theta_k) = \sum_{k=1}^{g} p_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\theta_k^{jh})^{x^{jh}}$$

  - Paramètre $\Phi = (\mathbf{p}, \theta)$:
    * Proportions $\mathbf{p} = (p_1, \ldots, p_g)$
    * Paramètres de chaque composant $\theta = (\theta_1, \ldots, \theta_g)$

Problème :

- Estimer le paramètre
$$\Phi = (\mathbf{p}, \theta)$$
- Eventuellement, estimer la classe d'origine
- Solution : maximiser la vraisemblance
- Nombre de paramètres :
  - Classes latentes : $(g-1) + g * \sum_j (m_j - 1)$
  - Modèle complet : $\prod_j m_j$
  - pour g=5 et d=10 et $m_j = 4$: 154 et environ 10^6

Vraisemblances

- Vraisemblance

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

$$= \sum_{i=1}^{n} \log\{\sum_{k=1}^{g} p_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\theta_k^{jh})^{x_i^{jh}}\}$$

- Vraisemblance complétée

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

Algorithme EM

- Étape 0 : solution initiale $\Phi^{(0)} = (\mathbf{p}^{(0)}, \theta^{(0)})$
- 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 M :

  - $p_k^{(c+1)} = \frac{\sum_i t_{ik}}{n}$
  - $(\theta_k^{jh})^{(c+1)} = \frac{\sum_{i=1}^{n} t_{ik} x_i^{jh}}{\sum_{i=1}^{n} t_{ik}}$

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 : $\mathbf{z}^{(c+1)} = \mathrm{MAP}(t_{ik})$:
  Les t_{ki} sont remplacées par des 1 ou des 0
- Étape M : maximisation de $L_C(\mathbf{z}^{(c+1)}, .)$

  – $p_k^{(c+1)} = \frac{\# z_k}{n}$
  – $(\theta_k^{jh})^{(c+1)} = \frac{\sum_{i=1}^{n} t_{ik} x_i^{jh}}{\# z_k} =$ % de modalité h

Modèle parcimonieux : reparamétrisation

Si h^* est la modalité de plus grande probabilité :

$$f(\mathbf{x};\Phi) = \sum_{k=1}^{g} p_k \prod_{j=1}^{d} \left((1-\varepsilon_k^{jh^*})^{1-d(x^j,a_k^j)} \prod_{h \neq h^*} (\varepsilon_k^{jh})^{|x^{jh}-a_k^{jh}|} \right)$$

- Centre $\mathbf{a}_k^j = (a_k^{j1}, \ldots, a_k^{jm_j})$ où $a_k^{jh} = \begin{cases} 1 \text{ si } h = h^* \\ 0 \text{ sinon} \end{cases}$
- Dispersions $\varepsilon_k^j = (\varepsilon_k^{j1}, \ldots, \varepsilon_k^{jm_j})$ où $\varepsilon_k^{jh} = \begin{cases} 1-\theta_k^{jh} \text{ si } h = h^* \\ \theta_k^{jh} \text{ sinon} \end{cases}$
- $d(x^j, a_k^j) = 0$ si x^j et a_k^j prennent la même modalité et 1 sinon
- Exemple : $\theta_k^j = (0.7, 0.2, 0.1)$: $\mathbf{a}_k^j = (1, 0, 0)$ et $\varepsilon_k^j = (0.3, 0.2, 0.1)$

- Modèle initial : ε_k^{jh}
- Modèle proposé en discrimination :

$$\begin{cases} \varepsilon_k^{jh^*} = \varepsilon_k^j \\ \varepsilon_k^{jh} = (1-\varepsilon_k^j)/(m_j-1) \text{ pour } h \neq h^* \end{cases}$$

- Exemple : $\mathbf{a}_k^j = (1,0,0)$ et $\varepsilon_k^j = (0.2, 0.1, 0.1)$ ou $\varepsilon_k^j = (0.3, 0.15, 0.15)$
- Modèle plus parcimonieux :

$$\begin{array}{lll} \varepsilon_k : & \varepsilon_k^{jh^*} = \varepsilon_k & \forall j \\ \varepsilon^j : & \varepsilon_k^{jh^*} = \varepsilon^j & \forall k \\ \varepsilon : & \varepsilon_k^{jh^*} = \varepsilon & \forall k, j \end{array}$$

Expression de la vraisemblance complétée

On se limite au cas le plus simple : modèle ε et $p_k = 1/g$

on a alors

$$L_C(\mathbf{z}, \Phi) = \log(\frac{\varepsilon}{1-\varepsilon}) \sum_{k=1}^{g} \sum_{i \in z_k} d(\mathbf{x}_i, \mathbf{a}_k) + nd \log(1-\varepsilon)$$

où

$$d(\mathbf{x}_i, \mathbf{a}_k) = \text{nb de modalités différentes entre } x_i \text{ et } a_k$$

Algorithme CEM correspondant

- Étape 0 : solution initiale $(\Phi^0, \mathbf{z}^0)$
- Étape E : les individus i sont rangés dans la classe k qui minimise

$$d(\mathbf{x}_i, \mathbf{a}_k)$$

- Étape M :
  - $p_k^{(c+1)} = \frac{\#z_k}{n}$
  - $(a_k^j)^{(c+1)} =$ modalité majoritaire de la classe k pour j

Autres approches

- Modèle binaire sur le tableau codage disjonctif complet
  - Codage binaire des variables qualitatives
  - À chaque modalité est associée sa variable binaire caractéristique
  - Exemple : 2 3 1 devient 010100001

Variables qualitatives ordinales

- Modèle parcimonieux adapté
- Modèle binaire sur le tableau additif
  - Codage binaire des variables qualitatives
  - Exemple : 2 3 1 devient 011111001

Conclusion

- Prendre les modèles les plus simples par classe
  - Modèle des classes latentes
  - Un seul paramètre λ
- Augmenter le nombre de composants si nécessaire

References

[1] J. Aitchison and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976.

[2] H. Bock. Loglinear models and entropy clustering methods for qualitative data. In W. Gaull and M. Schader, editors, *Classification as a tool of research*, pages 19–26. North-Holland, Amsterdam, 1986.

[3] H. Bock. Probabilistic aspects in cluster analysis. In O. Opitz, editor, *Conceptual and Numerical Analysis of Data*, pages 12–44. Springer-Verlag, Berlin, 1989.

[4] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176, 1991.

[5] M. Dang and G. Govaert. Fuzzy clustering of spatial binary data. *Kybernetika*, 34(4):393–398, 1998.

[6] B. Everitt. *An Introduction to Latent Variable Models*. Chapman and Hall, London, 1984.

[7] L. A. Goodman. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

[8] G. Govaert. Classification binaire et modèles. *Revue de Statistique Appliquée*, XXXVIII(1):67–81, 1990.

[9] G. Govaert and M. Nadif. Comparison of the mixture and the classification maximum likelihood in cluster analysis when data are binary. *Computational Statistics and Data Analysis*, 23:65–81, 1996.

[10] P. F. Lazarfield and N. W. Henry. *Latent Structure Analysis*. Houghton Mifflin Company, Boston, 1968.

[11] M. Nadif and G. Govaert. Binary clustering with missing data. *Applied Stochastic Models and Data Analysis*, 10:59–71, 1993.

[12] M. Nadif and G. Govaert. Clustering for binary data and mixture models: Choice of the model. *Applied Stochastic Models and Data Analysis*, 13:269–278, 1998.

[13] M. Nadif and F. Marchetti. Classification de données qualitatives et modèles. *Revue de statistique appliquée*, 41(1):55–69, 1993.

[14] S. A. Stouffer and J. Toby. Role conflict and personality. *American Journal of Sociology*, 56:395–506, 1951.