Modèles de classification non supervisée avec données manquantes non au hasard

Fabien Laporte [1] & Christophe Biernacki [2] & Gilles Celeux [3] & Julie Josse [4]

[1] *Centre de Mathématiques Appliques, École Polytechnique, Paris, France fabien.laporte@polytechnique.edu*
[2] *Inria Lille, Université de Lille, CNRS, France christophe.biernacki@inria.fr*
[3] *Inria Saclay, France, gilles.celeux@inria.fr*
[4] *Centre de Mathématiques Appliquées, École Polytechnique, Inria Saclay XPOP, France julie.josse@polytechnique.edu*

Résumé. La difficulté de prise en compte des données manquantes est souvent contournée en supposant que leur occurrence est due au hasard. Dans cette communication, nous envisageons que l'absence de certaines données n'est pas due au hasard dans le contexte de la classification non supervisée et nous proposons des modèles logistiques pour traduire le fait que cette occurrence peut être associée à la classification cherchée. Nous privilégions différents modèles que nous estimons par le maximum de vraisemblance et nous analysons leurs caractéristiques au travers de leur application sur des données hospitalières.

Mots-clés. Modèle de mélange, modèle logistique, données manquantes non au hasard (MNAR), algorithmes EM et Stochastique EM

Abstract. Usually missing data are assumed to be missing at random. In this talk, we propose logistic models assuming that missing data are not missing at random, in the model-based clustering setting, and that the occurrence of missing data is related to the clustering. Different models are proposed and estimated through the maximum likelihood methodology. Their characteristics are analyzed through numerical experiments on Hospital data.

Keywords. Model-based Clustering, Logistic Model, Missing Not At Random (MNAR) Data, EM and Stochastic EM Algorithms.

1 Le cadre de la modélisation

Trois types de données manquantes ont été décrits par Little et Rubin (1986) : les données manquantes complétement au hasard (MCAR), si l'absence de données ne dépend pas des données, les données manquantes au hasard (MAR), si cette absence ne dépend pas des valeurs manquantes, et les données manquantes non au hasard (MNAR) si cette absence dépend des valeurs manquantes.
Lorsque les données manquantes sont MNAR, elles peuvent perturber les résulats d'une

inférence statistique. Nous plaçons l'inférence dans le cadre de la recherche d'une classification par un modèle de mélange. Nous explorons des modèles MNAR qui présupposent que l'occurence de données manquantes est liée à la classification recherchée. Nous restreignons notre présentation au cas des mélanges gaussiens, mais l'extension de cette recherche aux modèles de mélange multinomiaux multivariés pour des données qualitatives ou de mélange de Poisson pour des données de comptage n'induit pas de difficulté particulière.

2 Modèles MNAR pour un mélange gaussien

Soit $\mathbf{y}_1, \ldots, \mathbf{y}_n$, n vecteurs iid de $\mathbb{R}^d$ de loi

$$f(\mathbf{y}_i; \theta) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i | \mu_k, \Sigma_k)$$

où K est le nombre de classes, π_k est la proportion de la classe k, et $\phi(.|\mu, \Sigma)$ est la densité d'un loi gaussienne de moyenne μ et de variance Σ. On note par ailleurs $\mathbf{z}_i$ les vecteurs indicateurs inconnus d'appartenance de l'individu i à la classe k. Le vecteur de paramètre de ce modèle de mélage est $\theta = ((\pi_k, \mu_k, \Sigma_k), k = 1, \ldots, K)$. Par la suite, $\mathbf{y} = (y_{i,j})$ sera la matrice composée des vecteurs $\mathbf{y}_1, \ldots, \mathbf{y}_n$ et $\mathbf{z} = (z_{i,k})$ sera la matrice composée des vecteurs $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

On suppose que de (nombreuses) données manquantes sont présentes et on note $\mathbf{c} = (c_{i,j})$ la matrice des indicatrices des données manquantes : $c_{i,j} = 1$ si et seulement $y_{i,j}$, la valeur de la variable j pour l'individu i, est manquant. Nous allons considérer des modèles MNAR où la répartition $\mathbf{c}$ des données manquantes peut dépendre de la valeur manquante mais aussi des classes du modèle de mélange à travers une équation logistique. Sous sa forme la plus générale, ce modèle s'écrit, ψ dénotant le vecteur de paramètres associé,

$$\begin{aligned}
&\text{logit}(P(c_{i,j}|\mathbf{y}, \mathbf{z}; \psi)) = \alpha_0 + \alpha_j y_{i,j} + \sum_{\ell \neq j} \alpha_\ell y_{i,\ell} + \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{z_{i,k}=1\}} \\
&\psi = (\alpha_0, \ldots, \alpha_d, \beta_1, \ldots, \beta_K)
\end{aligned}$$

Ce modèle est sans doute trop complexe pour être utile et nous nous sommes focalisés sur

quelques cas particuliers remarquables

$$\text{logit}(P(c_{i,j}|\mathbf{y},\mathbf{z};\psi)) = \alpha_0 \text{ (MCAR)}$$
$$\psi = \alpha_0$$
$$\text{logit}(P(c_{i,j}|\mathbf{y},\mathbf{z};\psi)) = \alpha_0 + \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{z_{i,k}=1\}} \text{ (MNAR}\mathbf{z}\text{)}$$
$$\psi = (\alpha_0, \beta_1, \ldots, \beta_K)$$
$$\text{logit}(P(c_{i,j}|\mathbf{y},\mathbf{z};\psi)) = \alpha_0 + \alpha_j y_{i,j} + \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{z_{i,k}=1\}} \text{ (MNAR}\mathbf{yz}\text{)}$$
$$\psi = (\alpha_0, \ldots, \alpha_d, \beta_1, \ldots, \beta_K)$$

Utilisant la nomenclature de Little et Rubin (1986), le premier de ces modèles (MCAR) est un modèle où les données manquent complétement au hasard. Le modèle (MNAR$\mathbf{z}$) est un modèle assez pauvre où les données manquantes n'interviennent que sur l'effectif des classes et le modèle MNAR$\mathbf{yz}$ semble le plus intéressant, car il allie un effet des classes avec un effet des différentes variables, tout en restant assez parcimonieux.

3 Estimation des modèles

L'estimation conjointe des paramètres θ et ψ par le maximum de vraisemblance peut se faire en principe à l'aide de l'algorithme EM de Dempster *et al.* (1977). Cet algorithme conduit effectivement à des formules explicites pour les modèles MCAR (pour lequel il est connu depuis longtemps, voir par exemple Hunt et Jorgensen, 2003) et MNAR$\mathbf{z}$.

En revanche, pour le modèle MNAR$\mathbf{yz}$, l'étape E de l'algorithme EM qui comporte l'actualisation des probabilités conditionnelles d'appartenance des individus aux classes sachant les données observées et la valeur courante des paramètres fait intervenir l'intégrale d'une densité gaussienne multipliée par une fonction logistique. Cette intégrale ne possède pas de formule exacte (cf. Pirjol (2013)) et est très difficile à calculer. Aussi, pour ce modèle, nous avons recours à l'algorithme Stochastique EM (SEM) (Celeux et Debolt,1985) dont le principe consiste à adjoindre aux étapes E et M de EM, une étape stochastique S de simulation des données et des labels manquants selon leur loi conditionnelle sachant les données observées et la valeur courante des paramètres.

4 Expérimentations

Pour analyser les performances de ces modèles, nous avons utilisé un jeu de données des hopitaux. Il comporte 5146 individus et 7 variables quantitatives avec un nombre important de données manquantes dont on peut penser qu'elles sont MNAR. Les analyses

que nous avons réalisées nous permettent de mesurer certaines qualités et certaines limites de ces modèles logistiques. Ils seront également illustrés sur des données simulées.

Bibliographie

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statitics Quaterly*, 2, 73-82.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39, 1-38.

Hunt, L. et Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis* 41, 429-440.

Little, R. J., et Rubin, D. B. (1986). *Statistical analysis with missing data*. Wiley.

Pirjol, D. (2013). The logistic-normal integral and its generalizations. *Journal of Computational and Applied Mathematics*, 237, 460-469.