Généralisation de l'algorithme Largest Gaps pour le modèle des blocs latents non-paramétrique

Vincent Brault [1] & Antoine Channarond [2] & Valérie Robert [3]

[1] *Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France*
CNRS, LJK, F-38000 Grenoble, France
vincent.brault@univ-grenoble-alpes.fr
[2] *UMR6085 CNRS, Laboratoire de Mathématiques Raphaël Salem, Université de Rouen Normandie, 76800 Saint-Étienne-du-Rouvray, France*
antoine.channarond@univ-rouen.fr
[3] *Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, F-91405 Orsay, France.*
INRIA Saclay Ile-de-France Projet Select, Université Paris-Sud, F-91405 Orsay, France.
Inserm UMR 1181, Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), F-94807 Villejuif, France.
valerie.robert@math.u-psud.fr

Résumé. Le modèle des blocs latents définit une loi pour chaque croisement de classe d'objets et de classe de variables d'un tableau de données ; les cases sont supposées indépendantes conditionnellement aux blocs formés. Pour estimer les paramètres, la plupart des algorithmes sont très coûteux en temps de calcul. Brault et Channarond (2016) ont proposé d'adapter l'algorithme *Largest Gaps*, qui utilise uniquement les marginales, au modèle des blocs latents binaire et ont obtenu une procédure estimant tous les paramètres du modèle de façon consistante mais nécessitant un grand nombre d'observations. Dans cet exposé, nous étendons la procédure au cas de toute loi ayant un moment d'ordre deux en l'associant à une estimation des marginales par l'algorithme EM.

Mots-clés. Modèle des blocs latents, algorithme Largest Gaps, modèle de mélange, algorithme EM, statistique non-paramétrique.

Abstract. The latent block model assumes there exists a distribution for each crossing between an object cluster and a variable cluster of a data table ; the cells are supposed to be independent conditionally to the choice of these clusters. To estimate the model parameters, most of algorithms are time consuming. Brault and Channarond (2016) proposed to adapt the *Largest Gaps* algorithm which consists in using the margins. They thus obtained a procedure which estimates all the model parameters consistently but requires a large number of observations. In this talk, we will extend the procedure to the case of any distribution having a second order moment by using an EM algorithm estimation.

Keywords. Latent block model, Largest Gaps algorithm, Mixture model, EM algorithm, Nonparametric statistic.

1 Introduction

Depuis quelques années, les méthodes de *coclustering* sont de plus en plus étudiées dans des domaines variés tels que le marketing, la génétique ou encore la sociologie. Parmi ces méthodes, le modèle des blocs latents est une méthode de classification non-supervisée et simultanée des lignes et des colonnes d'une matrice basée sur un modèle probabiliste. Le but est, notamment, de retrouver les classes des lignes et des colonnes de la matrice.

Pour résoudre ce problème de classification dans le cadre paramétrique, Govaert et Nadif (2008) proposent un algorithme EM variationnel (VEM), Keribin et al. (2015) ont étudié un algorithme stochastique appelé SEM et proposent des algorithmes bayésiens. Toutefois, ces algorithmes sont très coûteux en terme de temps ; pour répondre à ce problème Brault et Channarond (2016) proposent une adaptation de l'algorithme *Largest Gaps* dont le principe est de n'utiliser que les marginales et montrent, dans le cas de données binaires, que l'algorithme fournissait des estimations consistantes du nombre de classes, des classes et des paramètres.

Dans cet exposé, nous généralisons cet algorithme au cas d'un modèle des blocs latents non-paramétrique. La seule condition que nous mettons est que les lois utilisées possèdent un moment d'ordre 2. Pour cela, nous reprenons le principe de l'algorithme *Largest Gaps* en étudiant les marginales à l'aide d'un algorithme EM unidimensionnel.

2 Modèle

Soit $\mathbf{x} = (x_{ij})_{i=1,\dots,n;j=1,\dots,d} \in \mathbb{R}^{n\times d}$ une matrice de données réelles de dimension $n \times d$ mettant en relation n objets (observations) et d variables (attributs). L'objectif est d'opérer des permutations sur les lignes et sur les colonnes pour obtenir une réorganisation faisant apparaître des blocs contrastés. La partition $\mathbf{z}$ d'un échantillon $\{1,\dots,n\}$ en g classes est représentée par la matrice de classification $(z_{i,k})_{i=1,\dots,n;k=1,\dots,g}$ où $z_{i,k} = 1$ si i appartient à la classe k et 0 sinon. De façon similaire, la partition $\mathbf{w}$ d'un échantillon $\{1,\dots,d\}$ en m classes est représentée par la matrice de classification $(w_{j,\ell})_{j=1,\dots,d;\ell=1,\dots,m}$ où $w_{j,\ell} = 1$ si j appartient à la classe ℓ et 0 sinon. Les variables aléatoires sont notées en majuscule, la somme sur une ligne i d'une matrice $(a_{i,j})$ est représentée par $a_{i,+} = \sum_{j=1}^{d} a_{i,j}$ et sa moyenne par $\overline{a_{i,\cdot}}$.

Dans le modèle des blocs latents, nous faisons trois hypothèses :

1. Les distributions des variables latentes sont indépendantes $p(\mathbf{z},\mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$.
2. L'affection z_i de chaque ligne i à une classe est indépendante des autres affectations et suit une multinomiale $\mathcal{M}(1;\pi_1,\dots,\pi_g)$ (π_k est donc la probabilité pour une ligne d'appartenir à la classe k). De même, ρ_ℓ est la probabilité d'appartenance d'une colonne à la classe ℓ.
3. Connaissant le couple de partitions $(\mathbf{z},\mathbf{w})$, les variables $X_{i,j}$ sont indépendantes et

de loi $\mathcal{L}_{k,\ell}$ ne dépendant que du bloc dans lequel elles se trouvent :

$$X_{i,j}|z_{i,k}=1,\, w_{j,\ell}=1 \sim \mathcal{L}_{k,\ell}.$$

Dans cet exposé, nous supposons que toutes les lois $\mathcal{L}_{k,\ell}$ possèdent un moment d'ordre 2 ; c'est-à-dire que si $X \sim \mathcal{L}_{k,\ell}$ alors $\mathbb{E}[X^2] < +\infty$. Nous notons $\mu_{k,\ell}^{(r)} = \mathbb{E}[X^r]$ le moment d'ordre r de la loi.

3 Algorithme

Avant d'énoncer l'algorithme, nous expliquons le fondement théorique à l'origine.

3.1 Principe

Supposons que la ligne i appartienne à la classe k, alors, par la formule des espérances totales, nous avons pour tout entier $r \in \{1,2\}$ et toute colonne $j \in \{1,\dots,d\}$:

$$\mathbb{E}\left[X_{i,j}^r \middle|\, z_{i,k}=1\right] = \sum_{\ell=1}^{m} \mathbb{E}\left[X_{i,j}^r \middle|\, z_{i,k}=1, w_{j,\ell}=1\right] \mathbb{P}\left(w_{j,\ell}=1\right) = \sum_{\ell=1}^{m} \rho_\ell \mu_{k,\ell}^{(r)} =: \tau_k^{(r)}.$$

De plus, connaissant l'appartenance de la ligne i à la classe k les cases sont indépendantes. Ainsi, par le théorème de limite centrale, nous avons qu'asymptotiquement en d :

$$\overline{X_{i,\cdot}} \middle|\, z_{i,k}=1 \underset{+\infty}{\sim} \mathcal{N}\left(\tau_k^{(1)}, \frac{\tau_k^{(2)} - \left[\tau_k^{(1)}\right]^2}{d}\right).$$

Au final et par indépendance conditionnelle, la loi de la variable $\overline{X_{i,\cdot}}$ peut être approchée par un modèle de mélange de lois gaussiennes lorsque d est suffisamment grand.

3.2 Algorithme *Largest Gaps étendu*

L'algorithme *Largest Gaps étendu* est ainsi défini :

Algorithme Largest Gaps étendu :

Entrées : la matrice (x_{ij}), le nombre maximal de classes en ligne $g_{\max}$ et en colonne $m_{\max}$.

1. Pour tout $i \in \{1, \ldots, n\}$, calcul de $\overline{X_{i,\cdot}} = X_{i,+}/d$.
2. Pour tout $g \in \{1, \ldots, g_{\max}\}$, estimation à l'aide d'un algorithme *EM* gaussien des g classes.
3. Utilisation d'un critère de sélection de modèle pour le choix de g.

Même procédure pour les colonnes avec $m_{\max}$.
Renvoi des estimateurs g^{LG}, z^{LG}, m^{LG} et w^{LG}.

Remarque *: Si les cases $X_{i,j}$ appartiennent à un espace $\mathbb{R}^p$ avec $p \geq 2$, nous pouvons étendre la procédure en utilisant des gaussiennes multivariées.*

Remarque *: Dans certains cas, nous avons accès à la loi exacte. Par exemple, si les lois sont des Bernoulli, la loi de la somme est une loi binomiale. À ce moment là, nous montrons empiriquement qu'il est préférable d'utiliser la loi exacte pour l'estimation des classes à partir des marginales.*

4 Résultats théoriques

Dans cette partie, nous explicitons la complexité de notre algorithme et notre conjecture sur la consistance des estimateurs.

4.1 Consistance

En reprenant les démonstrations de l'article de Brault et Channarond (2016), nous montrons que les variables $\overline{X_{i\cdot}}$ d'une même classe k se concentrent autour de la moyenne $\tau_k^{(1)}$. Ainsi, nous avons la conjecture suivante :

Conjecture : Sous les hypothèses que :
- les paramètres $\tau_1^{(1)}, \ldots, \tau_{g^\star}^{(1)}$ (où $g^\star$ est le vrai nombre de classes en ligne) soient tous distincts,
- les paramètres π_k soient strictement positifs,
- le ratio $\log n/d$ tende vers 0 lorsque n et d tendent vers l'infini,

alors l'algorithme *Largest Gaps étendu* utilisé avec $g_{max} \geq g^\star$ et le critère ICL (*Integrated Complete Likelihood*) renvoie des estimateurs consistants du nombre de classes en ligne

et des partitions associées.

Remarque *: Nous avons la conjecture symétrique sur les colonnes.*

4.2 Complexité

En notant N_{Algo} la complexité maximale de l'algorithme utilisé pour estimer les classes, nous pouvons calculer la complexité de l'algorithme *Largest Gaps étendu*.

Théorème : La complexité de l'algorithme *Largest Gaps étendu* est :

$$\mathcal{O}\left(\max\left(nd, ng_{\max}^2 N_{Algo}, dm_{\max}^2 N_{Algo}\right)\right).$$

De plus, le calcul des $\overline{X_{i,\cdot}}$ et les g estimations des classes peuvent être parallélisées.

Remarque *: En reprenant la conjecture précédente, nous pouvons montrer que l'algorithme CEM renvoie un estimateur consistant des classes également. Si nous avons assez d'observations, nous pouvons donc l'utiliser afin de diminuer la complexité.*

5 Simulations

Pour évaluer la qualité des estimations de l'algorithme et comparer les résultats avec l'utilisation de l'algorithme EM exact ou approché par une loi gaussienne, nous avons fait un plan de simulations utilisant $g^\star = 5$, $m^\star = 4$ et des lois de Bernoulli $\mathcal{B}(\alpha_{k,\ell})$ avec

$$(\alpha_{k,\ell})_{\substack{k=1,\dots,g^\star \\ \ell=1,\dots,m^\star}} = \begin{pmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & 1-\varepsilon \end{pmatrix}$$

où $\varepsilon \in \{0.05, 0.25, 0.35, 0.45\}$. Nous avons pris des proportions équilibrées, à savoir :

$$\pi = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix} \text{ et } \rho = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}.$$

Enfin, nous avons fait varier le nombre de lignes n et de colonnes d de 20 en 20 en allant de 20 jusqu'à 800. Pour chaque cas, nous avons simulé 50 matrices et avons calculé l'erreur de classification des lignes (étudiée par Lomet (2012)) pour l'algorithme LG seul, couplé avec un algorithme EM binomial et avec un EM gaussien en admettant connaître le bon nombre de classes (voir figure 1).

Nous constatons que l'ajout de l'algorithme EM améliore l'estimation des labels et que l'approximation par une gaussienne ne marche qu'à partir d'un d assez grand dépendant de la difficulté.

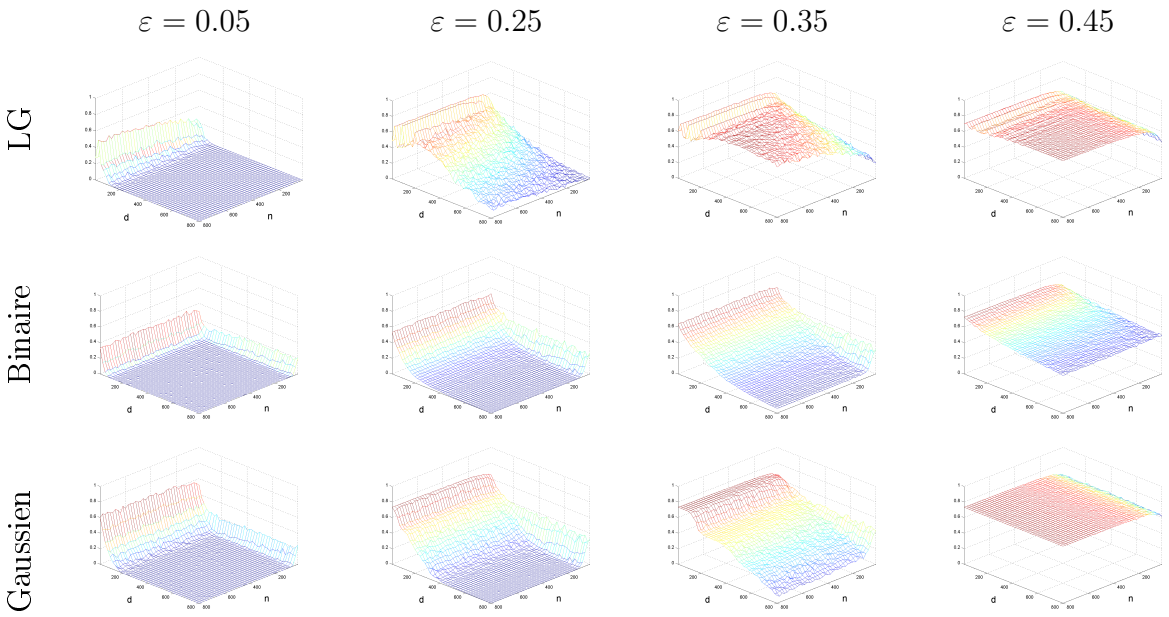

FIGURE 1 – Représentation des erreurs de classification des lignes en fonction de la difficulté (colonne) et de l'algorithme utilisé (ligne) : pour chaque graphique, nous avons représenté l'évolution en fonction du nombre de lignes n et de colonnes d.

6 Conclusions

Dans cet exposé, nous démontrerons la consistance des estimateurs et montrerons des résultats sur données simulées avec des lois différentes et sur des données réelles.

Bibliographie

[1] Brault, V. et Channarond, A. (2016). Fast and Consistent Algorithm for the Latent Block Model. *arXiv preprint arXiv :1610.09005*.
[2] 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), 1-38.
[3] Govaert, G. et Nadif, M. (2008). Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6), 3233-3245.
[4] Keribin, C., Brault, V., Celeux, G. et Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6), 1201-1216.
[5] Lomet, A. (2012). Sélection de modèle pour la classification croisée de données continues. *Doctoral dissertation*, Compiègne.