Mise en oeuvre de l'échantillonneur de *Gibbs* pour le modèle des blocs latents

Vincent Brault [1,2] & Gilles Celeux [2] & Christine Keribin[1,2]

[1] *Laboratoire de Mathématiques UMR* 8628*, Université Paris-Sud, F*-91405 *Orsay cedex*
[2] *INRIA Saclay Île de France Projet* SELECT*, Bat* 425*, Université Paris-Sud, F*-91405 *Orsay cedex*

Résumé. Les modèles de mélanges peuvent être utilisés pour résoudre le problème de la classification non supervisée simultanée d'un ensemble d'objets et d'un ensemble de variables. Le modèle des blocs latents définit une loi pour chaque croisement de classe d'objets et de classe de variables, et les observations sont supposées indépendantes conditionnellement au choix des classes d'objets et de variables. Mais il n'est pas possible de factoriser la loi jointe conditionnelle des labels rendant impossible le calcul de l'étape d'estimation de l'algorithme EM. Différents algorithmes existent pour contourner cette difficulté, notamment VEM, un EM variationnel, proposé par Govaert et Nadif (2008), l'algorithme SEM de Keribin et al (2010) ou encore d'un point de vue bayésien, l'algorithme *V-Bayes* proposé par Keribin et al (2012).
D'un point de vue théorique, l'échantillonneur de *Gibbs* (Keribin et al (2012)) permet de simuler la loi a posteriori exacte alors que d'autres algorithmes sont obligés de faire des approximations. D'un point de vue pratique, la question de l'atteinte de la stationnarité pour la chaîne générée en est un point délicat. Dans cet exposé, nous étudions la statistique de Brooks-Gelman (1998) comme critère d'arrêt pour le modèle des blocs latents et en proposons des améliorations pour diminuer le temps de convergence.

Mots-clés. Analyse de données et data mining - Échantillonnage de Gibbs - Statistique bayésienne - Classification croisée

Abstract. Mixture models can be used to deal with the simultaneous clustering of a set of objects and a set of variables. The latent block model defines a distribution for each combinaison of an object cluster and a variable cluster, and the data is supposed to be independent, given the object and the variable clusters. But the factorization of the joint distribution of the labels, conditionally to the observed data, is not tractable, and the E-step of the EM algorithm cannot be performed. To solve this problem, the variational EM has been proposed by Govaert and Nadif (2008), the SEM algorithm by Keribin and al (2010) and the *V-Bayes* algorithm by Keribin and al (2012).
In theory, the *Gibbs* sampler (Keribin et al (2012)) samples the exact a posteriori law while some algorithms use an approximation. In practice, the problem is to determine when the chain begins to be stationary. In this presentation, we study the Brooks-Gelman statistic

(1998) as stop criterion for the latent block model and propose some improvement to decrease the convergence period.

Keywords. Data mining - Gibbs sampling - Bayesian methods – Co-clustering

1 Introduction

Soit $\mathbf{x} = (x_{ij})_{i=1,\ldots,n \text{ et } j=1,\ldots,d} \in \{1,\ldots,r\}^{n\times d}$ une matrice de données catégorielles de dimension $n \times d$ mettant en relation n objets (observations) et d variables (attributs). Chaque case x_{ij} peut prendre des niveaux h non ordonnés allant de 1 à r. 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 $\{1,\ldots,n\}$ en g classes est représentée par la matrice de classification $(z_{ik}, i = 1,\ldots,n, k = 1,\ldots g)$ où $z_{ik} = 1$ si i appartient à la classe k et 0 sinon. De façon similaire, la partition $\mathbf{w}$ d'un échantillon $\{1,\ldots,d\}$ en m classes est représentée par la matrice de classification $(w_{j\ell}, j = 1,\ldots,d, \ell = 1,\ldots m)$ où $w_{j\ell} = 1$ si j appartient à la classe ℓ et 0 sinon. La probabilité d'appartenance pour une ligne i à la classe k sera notée π_k (et ρ_ℓ celle de l'appartenance d'une colonne à la classe ℓ). Les variables aléatoires sont notées en majuscule et la somme sur une ligne d'une matrice (a_{ij}) est représentée par $a_{.j} = \sum_i a_{ij}$. Enfin, nous notons (v_{ijh}) le tableau de dimension trois avec v_{ijh} valant 1 si la case x_{ij} vaut h et 0 sinon.
Pour résoudre ce problème de classification, Govaert et Nadif (2008) ont proposé un algorithme EM variationnel (VEM), Keribin et al (2010) ont étudié un algorithme stochastique appelé SEM et Keribin et al (2012) ont proposé des algorithmes bayésiens. En théorie, l'échantillonneur de *Gibbs* (Keribin et al (2012)) estime la probabilité a posteriori exacte $p(\mathbf{z}, \mathbf{w}, \theta|\mathbf{x})$ mais, en pratique, l'évaluation du moment où la chaîne a atteint la stationnarité s'avère difficile. Dans cet exposé, nous étudions différents critères d'arrêt basés sur la statistique de Brooks-Gelman (1998) et les comparerons en terme de qualité de classification et de temps de convergence.

2 Présentation du modèle des blocs latents

Chaque coefficient x_{ij} de la matrice $\mathbf{x}$ est le résultat du tirage d'une variable aléatoire X_{ij}. Dès que $\mathbf{z}$ et $\mathbf{w}$ sont fixés, la densité conditionnelle de la variable X_{ij} appartenant au bloc (k, ℓ) est $\varphi(.; \alpha_{k\ell})$. Comme dans l'analyse en classes latentes, nous supposons l'indépendance conditionnelle des $n \times d$ variables X_{ij} sachant le couple $(\mathbf{z}, \mathbf{w})$:

$$f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \theta) = \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}}.$$

Le modèle des blocs latents peut être défini comme un modèle de mélange

$$f(\mathbf{x};\theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z};\theta)p(\mathbf{w};\theta)f(\mathbf{x}|\mathbf{z},\mathbf{w};\theta)$$

où $\mathcal{Z}$ et $\mathcal{W}$ représentent les ensembles de toutes les affectations possibles $\mathbf{z}$ de $\{1,\ldots,n\}$ et $\mathbf{w}$ de $\{1,\ldots,d\}$.

Dans le cas des données catégorielles, nous définissons le paramètre $\theta = (\pi, \rho, \alpha_{k\ell}; k = 1,\ldots g, \ell = 1,\ldots,m)$, où $\pi = (\pi_1,\ldots,\pi_g)$ et $\rho = (\rho_1,\ldots,\rho_m)$, pour obtenir le modèle des blocs latents catégoriels :

$$f(\mathbf{x};\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

où $\alpha_{k\ell} = \left(\alpha_{k\ell}^h\right)_{h=1,\ldots,r} \in [0,1]^r$ avec $\sum_{h=1}^r \alpha_{k\ell}^h = 1$ et $\varphi(x_{ij};\alpha_{k\ell}) = \prod_{h=1}^r \left(\alpha_{k\ell}^h\right)^{v_{ijh}}$.
L'impossibilité de factoriser la loi jointe empêche le calcul numérique de la logvraisemblance et le calcul des lois conditionnelles nécessaires à l'algorithme EM. Différents algorithmes ont donc recours à des approximations pour contourner ces difficultés.

3 Échantillonneur de *Gibbs*

Dans un cadre bayésien (voir figure 1), le principe de l'échantillonneur de *Gibbs* est l'obtention d'une chaîne de Markov de loi stationnaire la loi a posteriori exacte $p(\mathbf{z},\mathbf{w},\theta|\mathbf{x})$. Nous proposons les lois a priori suivantes (voir Keribin et al. (2013)) :

$$\pi \sim \mathcal{D}(4,\ldots,4), \quad \rho \sim \mathcal{D}(4,\ldots,4) \text{ et } \alpha_{k\ell} \sim \mathcal{D}(1,\ldots,1).$$

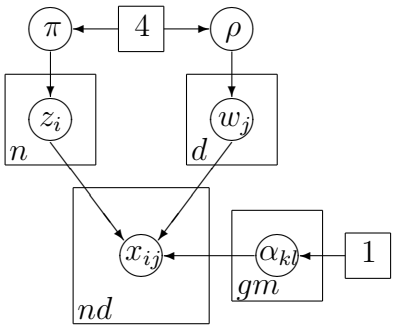

FIGURE 1 – Représentation schématique du modèle bayésien.

L'algorithme s'écrit :

Échantillonneur de *Gibbs* :

Itérations successives du schéma de *Gibbs* :

1. Simulation de $z^{(c+1)}$ suivant la loi $p(z|x, w^{(c)}; \theta^{(c)})$.
2. Simulation de $w^{(c+1)}$ suivant la loi $p(w|x, z^{(c+1)}; \theta^{(c)})$.
3. Simulation de $\pi^{(c+1)}$ suivant la loi $\pi|z^{(c+1)} \sim \mathcal{D}\left(z_{.1}^{(c+1)} + 4, \ldots, z_{.g}^{(c+1)} + 4\right)$.
4. Simulation de $\rho^{(c+1)}$ suivant la loi $\rho|w^{(c+1)} \sim \mathcal{D}\left(w_{.1}^{(c+1)} + 4, \ldots, w_{.m}^{(c+1)} + 4\right)$.
5. Simulation de $\alpha^{(c+1)}$ suivant la loi

$$\alpha_{k\ell}|x, z^{(c+1)}, w^{(c+1)} \sim \mathcal{D}\left(N^1_{k\ell;z,w}{}^{(c+1)} + 1, \ldots, N^r_{k\ell;z,w}{}^{(c+1)} + 1\right)$$

avec $N^h_{k\ell;z,w}{}^{(c+1)} = \sum_{ij} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} v_{ijh}$ le nombre de h dans le bloc (k, ℓ).

L'échantillonneur de *Gibbs* propose une méthode de simulation de la loi exacte a posteriori, celle-ci étant indépendante de l'initialisation choisie. En revanche, il nécessite de trouver un critère pour déterminer l'atteinte de la stationnarité, c'est-à-dire le moment à partir duquel la chaîne est simulée suivant la loi a posteriori visée.

4 Critère d'atteinte de la stationnarité

Pour savoir quand la chaîne générée par l'échantillonneur de *Gibbs* a atteint la stationnarité, nous proposons d'utiliser la statistique de Brooks-Gelman (voir Brooks et Gelman (1998)). Pour chaque composante de chaque paramètre (noté ξ), nous simulons τ chaînes en parallèle de longueur M et nous calculons, après un temps de chauffe, la statistique de la manière suivante :

1. Pour chaque chaîne $\xi_\iota = \left\{\xi_\iota^1, \ldots, \xi_\iota^M\right\}$, calcul de la différence δ_ι entre les quantiles empiriques de niveau 97.5% et 2.5%.
2. Calcul de la différence Δ entre les quantiles empiriques de niveau 97.5% et 2.5% pour l'échantillon complet $\{\xi_1, \ldots, \xi_\tau\}$.
3. Estimation de la statistique de Brooks-Gelman :

$$\widehat{R}_{BG} = \frac{\Delta}{\overline{\delta}}$$

où $\overline{\delta} = \frac{1}{\tau} \sum_{\iota=1}^{\tau} \delta_\iota$ est la moyenne empirique des δ_ι.

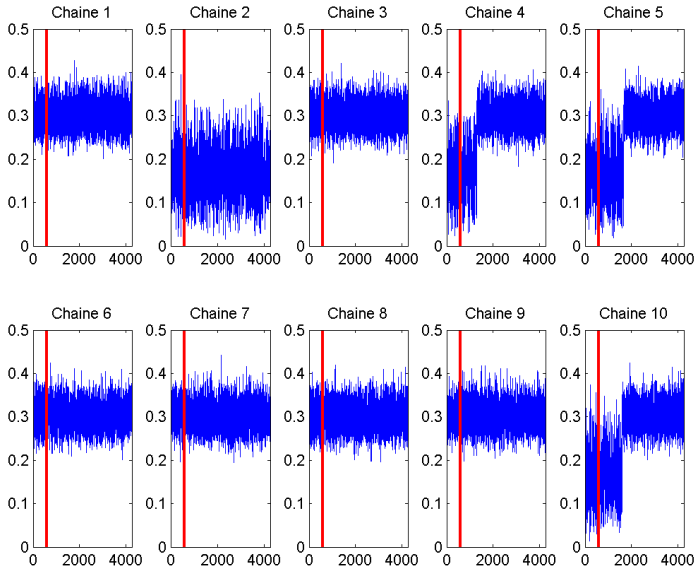

FIGURE 2 – Affichage des trajectoires de 10 chaînes pour l'un des paramètre π_k. Avant la ligne rouge se trouvent 20% des itérations.

Ce calcul est effectué à chaque fois que les τ chaînes possèdent un multiple de M itérations. Si $\widehat{R}_{BG}$ est inférieure à 1.2, l'algorithme s'arrête.

Pour l'échantillonneur de *Gibbs* proposé dans le cadre du modèle des blocs latents, nous ajoutons deux améliorations. Sur la figure 2, les chaînes 1, 3, 6, 7, 8 et 10 ont atteint la stationnarité. En revanche, la chaîne 2 semble être prise dans un maximum local ; tant qu'elle n'en sera pas sortie, la chaîne globale ne sera pas stationnaire. De même, les chaînes 4, 5 et 10 l'ont quitté après un nombre d'itérations compris entre 1500 et 2000 et il faudra attendre un grand nombre d'itérations pour contrecarrer l'effet négatif sur la convergence de ces premières itérations.

Pour éliminer ces problèmes, au moment du calcul de $\widehat{R}_{BG}$ et si celle ci est trop grande, nous proposons :

- de calculer, pour chaque ι, la statistique $\widehat{R}_{BG}^{-\iota}$ sur le même principe mais en enlevant la chaîne ι,
- de calculer la statistique $\widehat{R}_{BG}^{-20\%}$ en enlevant pour chaque chaîne les $0, 2M$ premières itérations,
- de calculer, pour chaque ι, la statistique $\widehat{R}_{BG}^{-(\iota,20\%)}$ en couplant les deux points précédents.

En contrepartie, nous comparons ces statistique à la valeur 1.05 afin d'assurer que l'échantillonneur de *Gibbs* ait bien convergé. Malgré cette restriction, nous montrerons que les

améliorations permettent de diminuer le temps de convergence sans dégrader la qualité des résultats.

5 Conclusion

Jusqu'à présent, pour le modèle des blocs latents, l'échantillonneur de *Gibbs* a été utilisé avec un nombre d'itérations fixé à l'avance. Nous comparerons cette procédure avec celle utilisant la statistique de Brooks-Gelman et avec les différentes améliorations proposées en terme de qualité de classement et de temps sur des données simulées et réelles.

Bibliographie

[1] Govaert, G. et Nadif M. (2008) Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52, 3233–3245.
[2] Keribin, C., Celeux, G. et Govaert, G. (2010) Estimation d'un modèle à blocs latent par l'algorithme SEM. *42° journées de Statistique*, SFdS, Marseille, France, mai 2010.
[3] Keribin, C., Brault, V., Celeux, G. et Govaert, G. (2012) Estimation and Selection for the Latent Block Model on Categorical Data. *Rapport de recherche RR-8264*, INRIA.
[4] Brooks, S.P.Gelman, A. (1998) General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **15**, 434-455.
[5] Fu, S (2012) Inversion probabiliste bayésienne en analyse d'incertitude. *Thèse*, Université Paris-Sud 11, 52-58.