Normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de blocs latents

Mahendra Mariadassou, Vincent Brault, Christine Keribin

Normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de blocs latents

Mahendra Mariadassou [1] & Vincent Brault [2] & Christine Keribin [3,4]

[1] MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France
[2] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
[3] *Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, F-91405 Orsay, France.*
[4] *INRIA Saclay Île-de-France Projet* SELECT, *Université Paris-Sud,Université Paris-Saclay, F-91405 Orsay, France.*

Résumé. Le modèle de blocs latents est une méthode non supervisée de classification simultanée des n lignes et d colonnes d'une matrice basée sur un modèle probabiliste de mélange. Si les méthodes d'estimation sont maintenant bien maîtrisées, les résultats concernant l'asymptotique de l'estimateur du maximum de vraisemblance restent encore parcellaires. Sous certaines conditions de bornes sur les paramètres, et pour un régime asymptotique tel que $\log(d)/n$ et $\log(n)/d$ tendent vers 0 avec n et d, nous montrons l'équivalence asymptotique du rapport de vraisemblance observée avec celui de la vraisemblance complète. Cette équivalence permet de transférer les propriétés de normalité asymptotique de l'estimateur du maximum de vraisemblance du modèle complet à l'estimateur du maximum de vraisemblance.

Mots-clés. Modèle de blocs latents, consistance, asymptotique, maximum de vraisemblance, inégalités de concentration

Abstract. Latent Block Model (LBM) is a model-based method to cluster simultaneously the d columns and n rows of a data matrix. Parameter estimation in LBM is a difficult and multifaceted problem. Although various estimation strategies have been proposed and are now well understood empirically, theoretical guarantees about their asymptotic behaviour is rather sparse. We show here that under some mild conditions on the parameter space, and in an asymptotic regime where $\log(d)/n$ and $\log(n)/d$ go to 0 when n and d go to $+\infty$, (1) the maximum-likelihood estimate of the complete model (with known labels) is consistent and (2) the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models. This equivalence allows us to transfer the asymptotic consistency to the maximum likelihood estimate under the observed model.

Keywords. Latent Block Model, Asymptotic Consistency, Maximum Likelihood Estimate, Concentration Inequality

1 Introduction

Le modèle de blocs latents (LBM, *Latent Block Model*) est une méthode non supervisée de classification simultanée des lignes et colonnes d'une matrice (*coclustering*), basée sur un modèle probabiliste. Les méthodes de coclustering connaissent un intérêt croissant depuis quelques années, notamment grâce à de nombreuses applications dans des domaines très variés tels que l'analyse textuelle, la génomique ou les systèmes de recommandation.

Nous observons une matrice $X = (x_{ij})$ de n lignes et d colonnes et supposons que les lignes sont partitionnées en g classes de lignes et les colonnes en m classes de colonnes. L'appartenance de chaque ligne (resp. colonne) à sa classe en ligne (resp. colonne) n'est pas connue et doit être déterminée. Une fois ces appartenances déterminées, le ré-ordonnancement des lignes et colonnes suivant leurs classes respectives fait apparaître des blocs d'intérieur homogène et contrastés entre eux, amenant à une représentation particulièrement parcimonieuse des données.

Le LBM s'applique aux donnéees binaires (Govaert et Nadif, 2008), gaussiennes (Lomet, 2012), catégorielles (Keribin et al., 2014) et aux donnéees de comptage (Govaert et Nadif, 2013). De part la structure complexe de dépendance, ni la vraisemblance, ni l'étape E (calcul de la loi des labels conditionnellement aux observations) de l'algorithme EM ne peuvent être calculées numériquement. L'estimation peut cependant être effectuée, soit par approximation variationnelle (menant à une valeur approchée de l'estimation du maximum de vraisemblance), soit par approche bayésienne (algorithmes VBayes ou échantillonneur de Gibbs), la recommandation de Keribin et al. (2014) étant de combiner ces deux derniers algorithmes.

Si l'estimation a trouvé des solutions, l'étude des propriétés théoriques des estimateurs (consistance, loi asymptotique) est restée pour l'instant ouverte. Des résultats partiels existent cependant pour LBM et pour SBM (*Stochastic Block Model*), cas particulier de LBM où les données consistent en un graphe aléatoire encodé par sa matrice d'adjacence (les lignes et les colonnes représentent les mêmes individus, avec les mêmes labels pour les lignes et les colonnes). Celisse et al (2012) ont montré pour SBM que, sous la vraie valeur du paramètre, la loi des labels conditionnellement aux observations tend vers un Dirac des vrais labels (Theorem 3). Cette convergence est de plus valide sous la valeur estimée du paramètre si l'estimateur converge à une vitesse d'au moins n^{-1} vers la vraie valeur, où n est le nombre de noeuds du graphe (Proposition 3.8). Cette hypothèse n'est pas anodine, et il n'est pas établi qu'un tel estimateur existe sauf dans certains cas particuliers (Ambroise et Mattias, 2011 par exemple). Mariadassou et Matias (2015) ont présenté un cadre unifiant SBM et LBM pour des observations de familles exponentielles, et montré la convergence de la loi conditionnelle des labels pour toute valeur du paramètre dans un voisinage de la vraie valeur. Bickel et al (2013) ont prouvé la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle SBM binaire. Rompant avec la vision des précédents auteurs, ils ont d'abord étudié le comportement asymptotique de l'estimateur du maximum de vraisemblance dans le modèle

complet (observations et labels) qui est plus simple à manipuler, puis ont prouvé que la vraisemblance complète et la vraisemblance des observations avaient des comportements asymptotiques similaires, en utilisant en particulier une inégalité de Bernstein pour des observations bornées.

Nous étendons les résultats précédents au LBM pour des observations de familles exponentielles, en suivant le schéma de Bickel et al (2013). Nous commençons par préciser la définition du LBM, puis nous étudions l'estimateur du maximum de vraisemblance du modèle complet (observations et labels). Enfin, nous présentons le résultat principal permettant d'établir le comportement asymptotique de l'estimateur du maximum de vraisemblance.

2 Modèle

Le modèle LBM suppose une structure en blocs de données obtenus par le produit cartésien d'une partition des lignes par une partition des colonnes. Plus précisément,

- les labels des lignes $\mathbf{z}_i$, $i = 1, \ldots, n$, sont indépendants des labels des colonnes $\mathbf{w}_j$, $j = 1, \ldots, d$: $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$;
- les labels des lignes sont i.i.d.: $\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_g))$; de même, les labels des colonnes sont i.i.d.: $\mathbf{w}_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_m))$. $\boldsymbol{\pi}$ (resp. $\boldsymbol{\rho}$) représente les poids du mélange pour les lignes (resp. colonnes);
- conditionnellement aux blocs d'appartenance $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \times (\mathbf{w}_1, \ldots, \mathbf{w}_d)$, les observations X_{ij} sont indépendantes, de lois appartenant à la même famille paramétrique, de paramètre ne dépendant que du bloc considéré:

$$X_{ij}|z_{ik}w_{j\ell} = 1 \sim \varphi(., \alpha_{k\ell}).$$

Ainsi, le paramètre est $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$ et la log-vraisemblance des observations s'écrit

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(X; \boldsymbol{\theta}) = \log \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \left(\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}} \right) \tag{1}$$

où $\mathcal{Z} \times \mathcal{W}$ est l'ensemble de toutes les affectations possibles de $\mathbf{z}$ et $\mathbf{w}$.

Nous supposons que la densité φ appartient à une famille exponentielle régulière mise sous forme canonique et de paramètre unidimensionnel α :

$$\varphi(x, \alpha) = b(x) \exp(\alpha x - \psi(\alpha)), \tag{2}$$

où α appartient à l'espace $\mathcal{A}$, de telle sorte que $\varphi(\cdot, \alpha)$ est bien définie pour tout $\alpha \in \mathcal{A}$. Les propriétés classiques des familles exponentielles assurent que ψ est convexe, infiniment différentiable sur $\mathring{\mathcal{A}}$, que $(\psi')^{-1}$ est bien défini sur $\psi'(\mathring{\mathcal{A}})$ et que $\mathbb{E}[X_\alpha] = \psi'(\alpha)$ et $\mathbb{V}[X_\alpha] = \psi''(\alpha)$ quand $X_\alpha \sim \varphi(., \alpha)$.

Nous posons également les hypothèses suivantes sur l'espace des paramètres :

H_1 : Il existe une constante positive c et un compact C_α tels que

$$\boldsymbol{\Theta} \subset [c, 1-c]^g \times [c, 1-c]^m \times C_\alpha^{g \times m} \quad \text{et} \quad C_\alpha \subset \mathring{\mathcal{A}}.$$

H_2 : La fonction $\alpha \mapsto \varphi(\cdot, \alpha)$ est injective.

H_3 : Le vrai paramètre $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ est dans l'intérieur de $\boldsymbol{\Theta}$.

H_4 : $\boldsymbol{\alpha}^\star$ ne contient ni deux lignes identiques, ni deux colonnes identiques.

Ces hypothèses sont standard. L'hypothèse H_1 permet de garantir que les poids du mélange sont raisonnablement loin de 0 et 1 pour assurer l'existence de chaque groupe. Elle permet également d'assurer que $\boldsymbol{\alpha}$ n'est pas à la frontière de $\mathcal{A}$ et que ψ' est Lipschitz sur un voisinage de C_α. Les hypothèses H_2 à H_4 sont nécessaires à l'identifiabilité du modèle. Comme $\boldsymbol{\alpha}$ est restreint à un sous-ensemble borné de $\mathcal{A}$, la variance de X_α vérifie

$$\sup_{\alpha \in C_\alpha} \mathbb{V}(X_\alpha) = \bar{\sigma}^2 < +\infty \quad \text{et} \quad \inf_{\alpha \in C_\alpha} \mathbb{V}(X_\alpha) = \underline{\sigma}^2 > 0. \tag{3}$$

Les hypothèses posées sont en particulier vérifiées dans le cas des variables de Bernoulli, de Poisson d'espérance est bornée, et normales de variance connue et d'espérance bornée.

3 Étude du modèle complet

Suivant Bickel et al (2013), nous commençons par montrer les propriétés de l'estimateur du maximum de vraisemblance dans le modèle complet. Le logarithme de la vraisemblance du modèle complet (observations et labels) s'écrit

$$\mathcal{L}(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \log p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \varphi(x_{ij}; \alpha_{k\ell}).$$

L'estimateur $\widehat{\boldsymbol{\theta}}_{MC}$ s'en déduit aisément, ainsi que son comportement asymptotique :

$$\hat{\pi}_k = \widehat{\pi}_k(\mathbf{z}) = \frac{z_{+k}}{n}; \quad \hat{\rho}_\ell = \widehat{\rho}_\ell(\mathbf{w}) = \frac{w_{+\ell}}{d}; \quad \hat{x}_{k\ell}(\mathbf{z}, \mathbf{w}) = \frac{\sum_{ij} x_{ij} z_{ik} w_{j\ell}}{z_{+k} w_{+\ell}}$$

$$\hat{\alpha}_{k\ell} = \widehat{\alpha}_{k\ell}(\mathbf{z}, \mathbf{w}) = (\psi')^{-1}\left(\hat{x}_{k\ell}(\mathbf{z}, \mathbf{w})\right)$$

Propriété 3.1 *Sous les hypothèses H_1 à H_4, les matrices $\Sigma_{\boldsymbol{\pi}^\star} = \mathrm{Diag}(\boldsymbol{\pi}^\star) - \boldsymbol{\pi}^\star (\boldsymbol{\pi}^\star)^T$, $\Sigma_{\boldsymbol{\rho}^\star} = \mathrm{Diag}(\boldsymbol{\rho}^\star) - \boldsymbol{\rho}^\star (\boldsymbol{\rho}^\star)^T$ sont semi-definies positives, de rang $g-1$ et $m-1$, et $\widehat{\boldsymbol{\pi}}$ et $\widehat{\boldsymbol{\rho}}$ sont asymptotiquement normaux:*

$$\sqrt{n}\left(\hat{\boldsymbol{\pi}}\left(\mathbf{z}^\star\right) - \boldsymbol{\pi}^\star\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\pi}^\star}) \quad et \quad \sqrt{d}\left(\hat{\boldsymbol{\rho}}\left(\mathbf{w}^\star\right) - \boldsymbol{\rho}^\star\right) \xrightarrow[d\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\rho}^\star}) \tag{4}$$

De façon similaire, soit $V(\boldsymbol{\alpha}^\star)$ la matrice définie par $V(\boldsymbol{\alpha}^\star)_{k\ell} = 1/\psi''(\alpha^\star_{k\ell})$ et $\Sigma_{\boldsymbol{\alpha}^\star} = \mathrm{Diag}^{-1}(\boldsymbol{\pi}^\star) V(\boldsymbol{\alpha}^\star) \mathrm{Diag}^{-1}(\boldsymbol{\rho}^\star)$. Alors:

$$\sqrt{nd}\left(\hat{\boldsymbol{\alpha}}\left(\mathbf{z}^\star, \mathbf{w}^\star\right) - \boldsymbol{\alpha}^\star\right) \xrightarrow[n,d\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\alpha}^\star}) \tag{5}$$

où la convergence se fait composante par composante et les composantes sont indépendantes.

Ainsi, pour tout k, ℓ, $\sqrt{nd}(\hat{\alpha}_{k\ell}\left(\mathbf{z}^\star, \mathbf{w}^\star\right) - \alpha^\star_{k\ell})$ tend en loi vers une gaussienne $\Sigma_{\boldsymbol{\alpha}^\star, k\ell}$ et est indépendant de tout autre $(\hat{\alpha}_{k'\ell'} - \alpha^\star_{k'\ell'})$.

4 Résultat principal

Avant d'énoncer le résultat principal permettant de transférer les propriétés de $\widehat{\boldsymbol{\theta}}_{MC}$ sur $\widehat{\boldsymbol{\theta}}_{MLE}$, l'estimateur du maximum de vraisemblance des données observées, nous définissons l'équivalence de deux paramètres.

Definition 4.1 (Équivalence) *Deux paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\theta}'$ sont équivalents, s'ils sont égaux à permutations près, c'est à dire s'il existe deux permutations s et t telles que $(\boldsymbol{\alpha}^{s,t}, \mathbf{z}^s, \mathbf{w}^t) = (\boldsymbol{\alpha}', \mathbf{z}', \mathbf{w}')$. On note alors $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$.*

Théorème 4.2 (Équivalence asymptotique) *Sous les hypothèses H_1 à H_4, et pour un régime tel que $\log(d)/n$ et $\log(n)/d$ tendent vers 0 avec n et d, les rapports de vraisemblance observée et de vraisemblance complète sont asymptotiquement équivalents :*

$$\frac{p(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}^\star)} = \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}')}{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} \left(1 + o_P(1)\right) + o_P(1)$$

où la convergence o_P est uniforme pour tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

La preuve repose sur l'étude du rapport de vraisemblance conditionnelle F_{nd} et du rapport de vraisemblance profilé $\Lambda(\mathbf{z}, \mathbf{w})$

$$F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \log \frac{p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)}; \quad \Lambda(\mathbf{z}, \mathbf{w}) = \max_{\boldsymbol{\theta}} F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}),$$

et distingue les cas (peu probables) de labels $(\mathbf{z}, \mathbf{w})$ très éloignés de la vraie configuration $(\mathbf{z}^*, \mathbf{w}^*)$ de ceux qui en sont proches. Des inégalités de concentration sont établies

en appliquant une inégalité de Bernstein à des variables non bornées de variance bornée (Massart 2007).

Les propriétés asymptotiques de l'estimateur du maximum de vraisemblance se déduisent du Théorème 4.2:

Corollaire 4.3 (Comportement asymptotique de $\widehat{\boldsymbol{\theta}}_{MLE}$) *Avec les notations de la Proposition 3.1 et en notant $\widehat{\boldsymbol{\theta}}_{MLE}$ l'estimateur du maximum de vraisemblance, il existe une permutation s de $\{1,\ldots,g\}$ et t de $\{1,\ldots,m\}$ telle que*

$$\begin{aligned} \hat{\boldsymbol{\pi}}\left(\mathbf{z}^{\star}\right) - \widehat{\boldsymbol{\pi}}^{s}_{MLE} = o\left(n^{-1/2}\right), \qquad & \hat{\boldsymbol{\rho}}\left(\mathbf{w}^{\star}\right) - \widehat{\boldsymbol{\rho}}^{t}_{MLE} = o\left(d^{-1/2}\right), \\ \hat{\boldsymbol{\alpha}}\left(\mathbf{z}^{\star},\mathbf{w}^{\star}\right) - \widehat{\boldsymbol{\alpha}}^{s,t}_{MLE} &= o\left((nd)^{-1/2}\right). \end{aligned}$$

Bibliographie

[1] Ambroise C. , Matias C. (2011), New consistent and asymptotically normal parameter estimates for random graph mixture models, *Journal of the Royal Statistical Society: Series B*

[2] Bickel P. , Choi D., Chang X. et Zhang H. (2013), Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels, *Ann. Stats*, 41(4):1922-1943

[3] Celisse A., Daudin J.-J. et Pierre L. (2012), Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model,*EJS*-DOI: 10.1214/154957804100000000.

[4] Govaert G. et Nadif M. (2008), Block clustering with Bernoulli mixture models: Comparison of different approaches, *Computational Statistics & Data Analysis*, 52, 3233–3245

[5] Govaert, G. et Nadif, M. (2013), Co-clustering. *John Wiley & Sons*.

[6] Keribin C., Brault V., Celeux G. et Govaert G (2014), Estimation and selection for the latent block model on categorical data, *STCO*, DOI 10.1007/s11222-014-9472-2

[7] Lomet A. (2012), Sélection de modèle pour la classification croisée de données continues, *Thèse de l'UTC, Compiègne*

[8] Massard P. (2007), Concentration Inequalities and Model Selection, Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003 *Springer*

[9] Mariadassou M. et Matias C. (2015), Convergence of the groups posterior distribution in latent or stochastic bloc model, *Bernoulli*, 21(1):537-573 DOI:10.3150/13-BEJ579