Un modèle statistique pour la pharmacovigilance

Valérie Robert, Gilles Celeux, Christine Keribin

Un modèle statistique pour la pharmacovigilance

Valérie Robert [(1,2,3)] & 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*, Bât.* 425*, Université Paris-Sud, F*-91405 *Orsay cedex*
[3] *Inserm UMR* 1181*, Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), F*-94807 *Villejuif, France*

Résumé. Les effets indésirables des médicaments sont le plus souvent découverts après l'autorisation de mise sur le marché de ces médicaments. La pharmacovigilance consiste alors à détecter le plus précocement possible l'existence d'associations entre médicaments et événements indésirables. Dans cette optique, des méthodes statistiques exploratoires (IC, Bate et al., 1998 ; GPS, Dumouchel, 1999...) sont développées depuis une vingtaine d'années. Cependant, ces méthodes sont limitées par l'utilisation de données agrégées (tableau de contingence), ce qui présume d'une homogénéité des individus à l'origine des notifications. Or il est raisonnable de supposer une certaine hétérogénéité dans la population étudiée. L'objectif est donc de proposer une alternative à ces méthodes intégrant cette dimension hétérogène du problème grâce à l'étude des données individuelles peu informatives, produisant des matrices creuses. Dans ce cadre, en adaptant le modèle des blocs latents (Govaert et Nadif, 2008), nous proposons un nouveau modèle statistique qui fournit une classification simultanée des lignes et des colonnes de deux tableaux de données binaires en leur imposant le même classement en ligne. Il permet alors d'établir des classes d'individus selon leur profil médicamenteux et des sous-groupes d'effets et de médicaments en interaction. Dans cet exposé, nous présenterons le modèle et montrerons la nouveauté de cette approche en pharmacovigilance. Nous donnerons des conditions suffisantes pour obtenir son identifiabilité et nous l'expérimenterons sur des matrices simulées creuses ou non.

Mots-clés. Pharmacovigilance – Algorithmes bayésiens – Modèles de mélange – Classification croisée – EM – Approximation variationnelle – Échantillonneur de Gibbs

Abstract. Adverse drug events are most often discovered after the marketing authorisation of these drugs. The pharmacovigilance system therefore aims at detecting as soon as possible potential associations between some drugs and adverse effects. From this standpoint, several statistical methods of automatic signal generation (IC, Bate and al., 1998 ; GPS, Dumouchel, 1999...) have been developed for over twenty years. Nevertheless, these explanatory methods suffer from limitations as they are based on aggregated data

(contingency table), which suppose some homogeneity in the individuals. But it is reasonable to believe that the studied population are heterogenous. The aim of this work is to propose an alternative to these methods by dealing with the heterogeneous dimension of the problem thanks to the study of the individual data which produce sparse matrices. Within this framework, a new approach is proposed in pharmacovigilance by developing a model adapted from the latent block model (Govaert and Nadif, 2008). It enables to co-cluster rows and columns of two binary tables by imposing the same row ranking. It also allows to hightlight subgroups of individuals sharing the same drug profile and subgroups of adverse effects and drugs with links. In this presentation, the model will be introduced first and sufficient conditions for its identifiability will be given. Then, the algorithms used for estimating the model will be developed and results on simulated sparse data will be presented.

Keywords. Pharmacovigilance – Bayesian Methods – Mixture Models – Co-clustering – EM – Variational Approximation – Gibbs Sampler

1 Introduction

Soient $x = (x_{ij})_{n\times J}$ et $y = (y_{ik})_{n\times K}$ deux matrices de données binaires, réalisations de deux variables aléatoires X et Y, telles que :

$$x_{ij} = \begin{cases} 1 & \text{si le médicament } j \text{ est présent} \\ & \text{dans la notification de l'individu } i, \\ 0 & \text{sinon,} \end{cases}$$

$$y_{ik} = \begin{cases} 1 & \text{si l'effet indésirable } k \text{ est présent} \\ & \text{dans la notification de l'individu } i, \\ 0 & \text{sinon.} \end{cases}$$

L'objectif est d'élaborer une classification simultanée des lignes et des colonnes de deux tableaux de données binaires en leur imposant le même classement en ligne afin d'obtenir un résumé faisant apparaître des blocs contrastés. Cette classification produit alors des classes d'individus selon leur profil médicamenteux ainsi que des sous-groupes d'effets et de médicaments en interaction. Dans l'exemple représenté en figure 1, les matrices de tailles respectives $(n, J) = (1000, 100)$ et $(n, K) = (1000, 100)$ peuvent être réduites en deux matrices de tailles respectives $(G, H) = (5, 4)$ et $(G, L) = (5, 3)$ et nous remarquons également que certains médicaments peuvent être mis en relation avec des effets indésirables.

Dans ce but, nous étendons le modèle des blocs latents (Govaert et Nadif, 2008) en construisant une partition des lignes et deux partitions des colonnes, l'une pour les colonnes de x et l'autre pour celles de y.

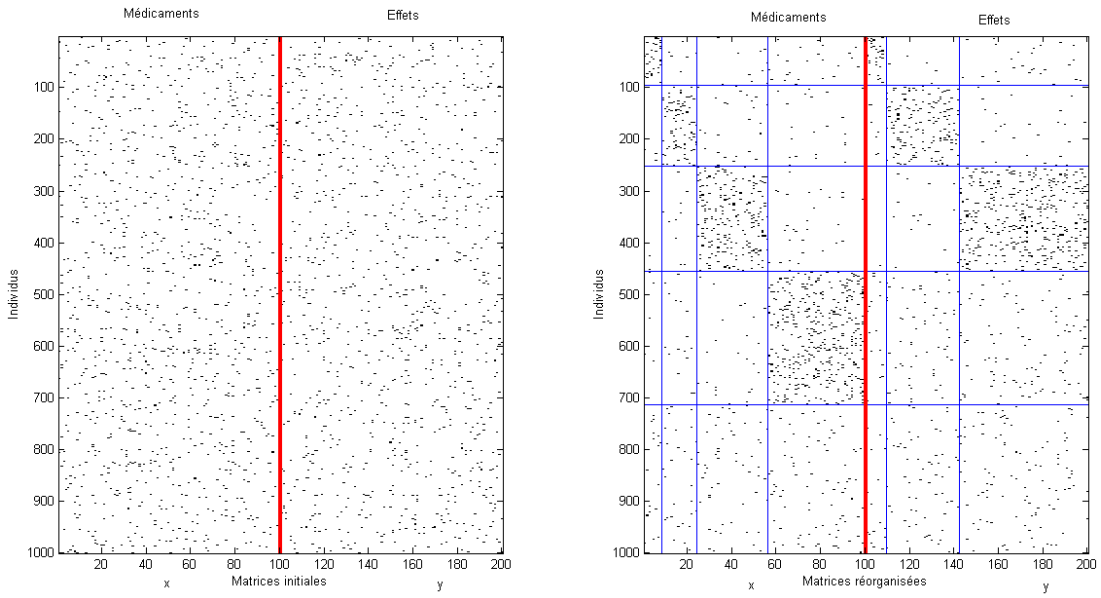

FIGURE 1 – *Matrices simulées x et y de données binaires (à gauche), réorganisées (à droite) avec la partition sur les effets, celle sur les médicaments et celle appariée sur les individus.*

2 Modèle des blocs latents multiple (MLBM)

2.1 Définition du modèle

Nous supposons que le nombre de classes en ligne et en colonne sont des paramètres fixes notés respectivement G, H et L. Dans ce cadre, nous faisons trois hypothèses :

(H_1) Les partitions en ligne $Z = (Z_1, \ldots, Z_G)$ et en colonne $V = (V_1, \ldots, V_H)$, $W = (W_1, \ldots, W_L)$ sont des variables latentes.

(H_2) L'indépendance a priori d'appartenance des classes en ligne et en colonne est supposée. Ainsi, les lignes ont toutes la même probabilité d'appartenir à la $g^{\text{ème}}$ classe en ligne : $\forall i \in \{1, \cdots, n\}$, $\mathbb{P}(Z_{ig} = 1) = \pi_g$. Nous notons $\pi = (\pi_1, \ldots, \pi_G)$ le vecteur des paramètres π_g et $\sum_{g=1}^{G} \pi_g = 1$.
De même, les colonnes de x (resp. y) ont toutes la même probabilité a priori ρ_h (resp. τ_ℓ) d'appartenir à la h ème classe (resp. ℓ ème classe).
Nous notons $\rho = (\rho_1, \ldots, \rho_H)$ et $\tau = (\tau_1, \ldots, \tau_L)$ avec $\sum_{h=1}^{H} \rho_h = 1$ et $\sum_{\ell=1}^{L} \tau_\ell = 1$.

(H_3) Enfin, les variables X et Y sont supposées indépendantes conditionnellement à la connaissance des appartenances aux classes en ligne et en colonne.

De plus, nous supposons que les variables X_{ij} et Y_{ik} suivent une loi conditionnelle de Bernoulli dont la densité est notée ϕ et dont les paramètres α_{gh} (resp. $\beta_{g\ell}$) dépendent du bloc (g,h) (resp. (g,ℓ)).

Nous obtenons alors le modèle des blocs latents multiple (MLBM) de Bernoulli qui peut être vu comme un modèle de mélange de densité :

$$\begin{aligned} & p(x,y;\theta) \\ &= \sum_{(z,v,w)\in\mathcal{Z}\times\mathcal{V}\times\mathcal{W}} p(z;\theta)p(v;\theta)p(w;\theta)p(x|z,v;\theta)p(y|z,w;\theta) \qquad (1) \\ &= \sum_{(z,v,w)\in\mathcal{Z}\times\mathcal{V}\times\mathcal{W}} \prod_{i,g} \pi_g^{z_{ig}} \prod_{j,h} \rho_h^{v_{jh}} \prod_{k,\ell} \tau_l^{w_{k\ell}} \prod_{i,j,g,h} \phi(x_{ij};\alpha_{gh})^{z_{ig}v_{jh}} \prod_{i,k,g,\ell} \phi(y_{ik};\beta_{g\ell})^{z_{ig}w_{k\ell}}, \end{aligned}$$

où $\mathcal{Z}$ (resp.$\mathcal{V}$, $\mathcal{W}$) est l'ensemble des partitions possibles des lignes (resp. des colonnes).

Nous résumons les paramètres des lois de Bernoulli par $\alpha = (\alpha_{gh})_{G\times H}$ et $\beta = (\beta_{g\ell})_{G\times L}$ et nous notons également $\theta = (\pi,\rho,\tau,\alpha,\beta)$.

2.2 Identifiabilité

Dans le cadre du modèle des blocs latents, Keribin et al. (2014) énonce des conditions suffisantes d'identifiabilité dans le cas de données binaires. En adaptant leur théorème, nous en déduisons des conditions suffisantes d'identifiabilité pour le modèle ci-dessus :

Conditions suffisantes d'identifiabilité. *Considérons le modèle MLBM et notons $A = (\alpha_{gh})$ et $B = (\beta_{g\ell})$. Définissons les conditions suivantes :*

- *C1 : pour tout $1 \le g \le G$, $\pi_g > 0$ et les coordonnées des vecteurs $\mu = A\rho$ (resp. $\nu = B\tau$) sont distinctes.*
- *C2 : Pour tout $1 \le h \le H$ et $1 \le \ell \le L$, $\rho_h > 0$ et $\tau_\ell > 0$ et les coordonnées du vecteur $\sigma = \pi' A$ (resp. $\upsilon = \pi' B$) sont distinctes où π' désigne la transposée de π.*

Sous ces conditions, le modèle MLBM est identifiable dès que $n \ge \max(2H-1, 2L-1)$ et $J + K \ge 2G - 1$.

3 Estimation des paramètres

3.1 Vraisemblance

Remarquons qu'il n'est pas possible de factoriser les termes dans (1) pour contourner le calcul de la somme sur chaque triplet. Dans ce cadre, l'algorithme EM (Dempster et al., 1977), classique pour estimer la vraisemblance d'un modèle de mélange n'est pas applicable et une approximation variationnelle est souvent proposée (Govaert et Nadif,

2008). À l'instar de Keribin et al. (2014), nous utilisons l'approche bayésienne qui permet d'éviter des solutions dégénérées. Nous couplons l'échantillonneur de Gibbs très peu sensible au problème d'initialisation, avec l'algorithme variationnel V-Bayes, ce qui fournit alors facilement une bonne approximation du mode a posteriori, utile pour estimer les partitions.

Ici, θ est supposé aléatoire. Comme les lois conjuguées des lois multinomiales sont les lois de Dirichlet et celles des lois de Bernoulli, les lois bêta, nous munissons les proportions de lois a priori :

$$\pi \sim \mathcal{D}(a, \ldots, a),\ \rho \sim \mathcal{D}(a, \ldots, a) \text{ et } \tau \sim \mathcal{D}(a, \ldots, a),$$

et nous supposons que les variables α_{gh} et $\beta_{g\ell}$ sont indépendantes et de même loi :

$$\alpha \sim \prod_{g,h} \mathcal{B}e(b, b) \text{ et } \beta \sim \prod_{g,\ell} \mathcal{B}e(b, b),$$

où a et b sont des hyperparamètres à déterminer.

3.2 Échantillonneur de Gibbs

L'échantillonneur de Gibbs consiste à générer une chaîne de Markov de loi stationnaire la loi a posteriori $p(z, v, w, \theta | x, y)$. L'algorithme est le suivant :

Gibbs : Itération successive du schéma de Gibbs après initialisation aléatoire :

1. Simulation de $z^{(c+1)}$ suivant la loi de densité $p(z|x, y, v^{(c)}, w^{(c)}, \theta^{(c)})$.
2. Simulation de $v^{(c+1)}$ suivant la loi de densité $p(v|x, y, z^{(c+1)}, \theta^{(c)})$.
3. Simulation de $w^{(c+1)}$ suivant la loi de densité $p(w|x, y, z^{(c+1)}, \theta^{(c)})$.
4. Simulation de $\theta^{(c+1)}$ suivant la loi de densité $p(\theta|x, y, z^{(c+1)}, v^{(c+1)}, w^{(c+1)})$.

A l'issue de l'échantillonneur, un post-traitement est effectué de façon à éviter le label switching. Le paramètre θ est alors estimé par la moyenne a posteriori après un temps de chauffe et les partitions sont estimées en affectant chaque ligne (ou colonne) à la classe obtenue majoritairement par $z_i^{(c)}$ (resp. $v_j^{(c)}, w_k^{(c)}$), simulés au cours des itérations. Nous utilisons alors ces résultats pour initialiser l'algorithme V-Bayes.

3.3 L'algorithme V-Bayes

L'algorithme V-Bayes cherche à maximiser *l'énergie libre* Q_B définie par :

$$\begin{aligned} Q_B(\theta) \ = \ & \sum_{i,g} s_{ig}^{(c)} \log \pi_g + \sum_{j,h} r_{jh}^{(c)} \log \rho_h + \sum_{i,j,g,h} s_{ig}^{(c)} r_{jh}^{(c)} \log \phi(x_{ij}; \alpha_{gh}) \\ & + \sum_{k,\ell} t_{k\ell}^{(c)} \log \tau_\ell + \sum_{i,k,g,\ell} s_{ig}^{(c)} t_{k\ell}^{(c)} \log \phi(y_{ik}; \beta_{g\ell}) \\ & - \sum_{i,g} s_{ig}^{(c)} \log s_{ig}^{(c)} - \sum_{j,h} r_{jh}^{(c)} \log r_{jh}^{(c)} - \sum_{k,\ell} t_{k\ell}^{(c)} \log t_{k\ell}^{(c)} + \log p(\theta), \end{aligned}$$

où $s_{ig}^{(c)} = \mathbb{P}(Z_{ig} = 1|x, y; \theta^{(c)})$, $r_{jh}^{(c)} = \mathbb{P}(V_{jh} = 1|x, y; \theta^{(c)})$, $t_{k\ell}^{(c)} = \mathbb{P}(W_{k\ell} = 1|x, y; \theta^{(c)})$,
$p(\theta)$ désigne la densité de la loi a priori de θ et $\theta^{(c)}$ est un paramètre courant.

Nous obtenons l'algorithme suivant :

V-Bayes : Itération des étapes suivantes jusqu'à n_{iter}, après initialisation :

1. Étape E : semblable à celle de l'algorithme VEM de Govaert et Nadif (2008) : maximisation alternée de s_{ig}, r_{jh} et $t_{k\ell}$.
2. Étape M-Bayes : calcul de $\theta^{(c+1)}$ tel que $\theta^{(c+1)} \in \operatorname{argmax} Q_B(\theta)$.

L'estimation des partitions se fait ensuite par la règle du maximum a posteriori (MAP) :

$$\widehat{z_i} \in \underset{g=1,\ldots,G}{\operatorname{argmax}}\, s_{ig}^{(n_{\text{iter}})}, \widehat{v_j} \in \underset{h=1,\ldots,H}{\operatorname{argmax}}\, r_{jh}^{(n_{\text{iter}})} \text{et } \widehat{w_k} \in \underset{\ell=1,\ldots,L}{\operatorname{argmax}}\, t_{k\ell}^{(n_{\text{iter}})}$$

4 Conclusion

Dans cet exposé, nous présenterons le modèle et donnerons des conditions suffisantes pour obtenir son identifiabilité. Ensuite, nous détaillerons les algorithmes utilisés pour sa mise en œuvre et nous l'expérimenterons sur des données simulées. Nous discuterons alors l'importance du choix des hyperparamètres a et b pour l'inférence bayésienne, lorsqu'on a affaire à des matrices creuses.

Bibliographie

[1] Bate, A. ; Lindquist, M. ; Edwards, I. ; Olsson, S. ; Orre, R. ; Lansner, A. ; et De Freitas, R. (1998) A Bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, 54, 315–321.
[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* 39 (1), 1–38.
[3] Dumouchel W. (1999) Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The american statistician*, 53, 177–190.
[4] Govaert, G. et Nadif, M. (2008) Block clustering with mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis*, 52, 3233–3245.
[5] Keribin, C. ; Brault V. ; Celeux, G. et Govaert, G. (2014) Estimation and selection for the latent block model on categorical data. *Statistics and computing*, http ://link.springer.com/article/10.1007/s11222-014-9472-2.