Modèle des blocs latents et sélection de modèles en pharmacovigilance

Valérie Robert [1,2,3] & Gilles Celeux [2] & Christine Keribin [1,2]
& Pascale Tubert-Bitter [3]

[1] *Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, F-*91405 *Orsay, France.*
[2] *INRIA Saclay Île-de-France Projet* SELECT*, Université Paris-Sud, F-*91405 *Orsay, France.*
[3] *Inserm UMR* 1181*, Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), F-*94807 *Villejuif, France.*

Résumé. La pharmacovigilance consiste à détecter le plus précocement possible l'existence d'associations entre médicaments et événements indésirables. Dans cette optique, des méthodes statistiques exploratoires de la base de notifications spontanées sont développées depuis une vingtaine d'années. Ces méthodes prennent en compte des données agrégées (tableau de contingence), ce qui suppose une homogénéité des individus à l'origine des notifications. Or il est raisonnable de supposer une certaine hétérogénéité dans la population étudiée. De plus, les matrices de données individuelles étant très grandes en taille, il est nécessaire de limiter le nombre de médicaments et d'effets indésirables auxquels on s'intéresse. Nous proposons alors l'utilisation du modèle des blocs latents sur tableau de contingence qui permettra de sélectionner des sous-groupes d'effets et de médicaments en interaction. Dans cet exposé, nous expliciterons le modèle des blocs latents sur tableau de contingence ainsi que les algorithmes utilisés pour l'estimation des paramètres du modèle. Ensuite, nous proposerons des critères de sélection de ce modèle et une procédure basée sur des initialisations récursives des algorithmes pour contourner le problème des données volumineuses. Enfin, nous présenterons les résultats obtenus sur les données simulées et réelles.

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

Abstract. The pharmacovigilance system aims at detecting as soon as possible potential associations between some drugs and adverse effects. From this standpoint, several explanatory methods of automatic signal generation have been developed for over twenty years. These methods 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 heterogeneous. Moreover, these matrices are so large that we have to select drugs and adverse effects beforehand. The latent block model on contingency table will be then considered in order to select subgroups of adverse effects and drugs with links. In this talk, this model will be introduced and the algorithms used for estimating the model will be developed. Then, model selection criteria will be proposed and a procedure based on recursive initializations of the algorithms will be presented to overcome the issue of large data and some experiments on simulated and real data will be finally shown.

Keywords. Pharmacovigilance – Bayesian Methods – Mixture Models – Co-clustering – EM – Variational Approximation – Model selection

1 Introduction

Soit $x = \{x_{ij}; i = 1, \ldots, n; j = 1, \ldots, d\}$, réalisation d'une variable aléatoire X telle que x_{ij} représente le nombre de notifications impliquant le médicament i et l'effet j.

L'objectif est d'élaborer une classification simultanée des lignes et des colonnes de ce tableau de contingence afin d'obtenir un résumé faisant apparaître des blocs contrastés. Cette classifi-

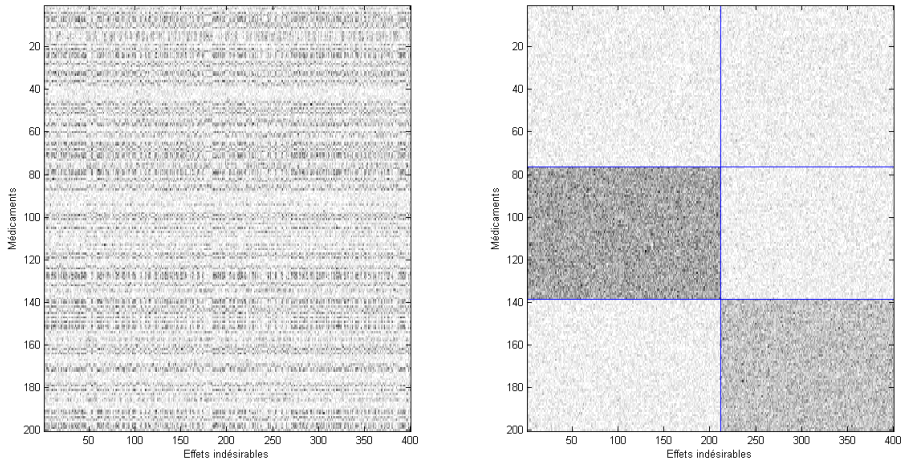

FIGURE 1 – *Matrice simulée x de données de comptage (à gauche), réorganisée (à droite) avec la partition sur les médicaments et celle sur les effets.*

cation croisée produit alors des sous-groupes d'effets et de médicaments en interaction. Dans l'exemple représenté en Figure 1, la matrice de taille $(n, d) = (200, 400)$ peut être réduite en une matrice $(G, H) = (3, 2)$ et nous remarquons également que certains médicaments peuvent être mis en relation avec des effets indésirables.

Dans ce but, nous étudions le modèle des blocs latents pour les observations de Poisson (Govaert et Nadif, 2010) et proposons une nouvelle procédure d'estimation des paramètres de ce modèle ainsi que des critères de sélection de ce modèle.

2 Modèle des blocs latents sur tableau de contingence

2.1 Définition du modèle

Pour introduire le modèle des blocs latents général (LBM, Govaert et Nadif, 2008), trois hypothèses sont nécessaires :

- Il existe une structure en blocs des données et ces blocs sont obtenus par le produit cartésien d'une partition des lignes en g composantes notée $z = (z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ et d'une partition des colonnes en m composantes notée $w = (w_{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots, m)$.
- Les labels des lignes et colonnes z et w sont indépendants a priori, c'est-à-dire

$$p(z, w) = p(z)p(w),$$

avec $p(z) = \prod_{i,k} \pi_k^{z_{ik}}$ et $p(w) = \prod_{j,\ell} \rho_\ell^{w_{j\ell}}$, où $(\pi_k = \mathbb{P}(z_{ik} = 1), k = 1, \ldots, g)$ et $(\rho_\ell = \mathbb{P}(w_{j\ell} = 1), \ell = 1, \ldots, m)$ sont les proportions des composantes en ligne et en colonne.

- Les variables aléatoires X_{ij} sont conditionnellement indépendantes sachant z et w. De plus, ces variables X_{ij} suivent une loi paramétrique notée ϕ.

Par conséquent, la densité marginale de x peut être vue comme une densité de mélange :

$$\begin{aligned} p(x;\theta) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} p(z;\theta)p(w;\theta)p(x|z,w;\theta) \\ &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \phi(x_{ij};\lambda_{k\ell})^{z_{ik}w_{j\ell}}, \end{aligned}$$

où $\mathcal{Z}$ et $\mathcal{W}$ représentent l'ensemble des partitions possibles pour les lignes et les colonnes, et $\theta = (\pi, \rho, \lambda)$ le paramètre du modèle à estimer.

Remarquons que le LBM s'applique aux données binaires (Govaert et Nadif, 2008), gaussiennes (Lomet, 2012), catégorielles (Keribin *et al.*, 2014) et aux données de comptage (Govaert et Nadif, 2013). Nous considérons cette dernière approche pour modéliser le tableau de contingence $x = \{x_{ij}; i = 1, \ldots, n; j = 1, \ldots, d\}$, où comme précédemment, x_{ij} représente le nombre de notifications impliquant le médicament i et l'effet j. Ainsi, la distribution conditionnelle $\phi(x_{ij};\lambda_{k\ell})$ de la variable X_{ij} sachant les labels z_{ik} et $w_{j\ell}$ est supposée être une loi de Poisson $\mathcal{P}(\mu_i\nu_j\lambda_{k\ell})$ où μ_i représente l'effet ligne (de la sorte, deux lignes proportionnelles seront mises dans le même bloc), ν_j représente l'effet colonne et $\lambda_{k\ell}$ représente l'interaction à l'intérieur du bloc $k\ell$. Cela conduit à estimer préalablement et de manière naturelle μ_i par la marginale $\sum_j x_{ij} = x_{i.}$ et ν_j par la marginale $\sum_i x_{ij} = x_{.j}$.

Ainsi, la densité par bloc s'écrit

$$\phi(x_{ij};\mu_i\nu_j\lambda_{k\ell}) = e^{-\mu_i\nu_j\lambda_{k\ell}} \frac{(\mu_i\nu_j\lambda_{k\ell})^{x_{ij}}}{x_{ij}!},$$

et la densité de ce modèle est alors

$$p(x;\theta) = \sum_{z,w} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_l^{w_{j\ell}} \prod_{i,j,k,\ell} \phi(x_{ij};\mu_i\nu_j\lambda_{k\ell}).$$

2.2 Estimation des paramètres

Dans cette partie, le nombre de classes g et m est fixé. L'algorithme EM, classique dans ce cadre, ne peut être mis en œuvre à cause de la structure de dépendance de la loi conditionnelle aux observations et une approche variationnelle est alors envisagée (Govaert et Nadif, 2008). De plus, pour éviter des solutions dégénérées, nous adaptons l'algorithme V-Bayes (Keribin *et al.*, 2014) qui est basé sur une inférence bayésienne (voir Figure 2) afin de fournir une estimation $\widehat{\theta}$ de θ.

Ainsi, θ est supposé ici aléatoire. Utilisant les lois conjuguées, les proportions de mélange sont munies de lois a priori de Dirichlet :

$$\pi \sim \mathcal{D}(a, .., a) \quad \text{et} \quad \rho \sim \mathcal{D}(a, .., a).$$

Nous choisissons le même hyperparamètre a pour toutes les distributions afin de ne favoriser aucune composante. Le paramètre λ est muni de la loi a priori Gamma :

$$\lambda_{k\ell}|\alpha, \beta \sim \Gamma(\alpha, \beta).$$

Le choix des hyperparamètres a, α, β s'effectuera selon les recommandations de Keribin *et al.* (2014). Par ailleurs, nous ne considérons pas ici μ_i et ν_j comme des variables aléatoires et ils sont alors estimés comme précédemment dit, par $x_{i.}$ et $x_{.j}$.

De plus, les conditions d'identifiabilité ci-dessous (Govaert et Nadif, 2013) :

$$\sum_i \mu_i = \sum_j \nu_j = \frac{1}{\sum_k \pi_k \lambda_{k\ell}} = \frac{1}{\sum_\ell \rho_\ell \lambda_{k\ell}} = \sum_{i,j} x_{ij},$$

assure les égalités suivantes,

$$\mathbb{E}(\sum_j x_{ij}) = \mu_i \quad \text{et} \quad \mathbb{E}(\sum_i x_{ij}) = \nu_j,$$

d'où l'estimation naturelle proposée pour μ_i et ν_j.

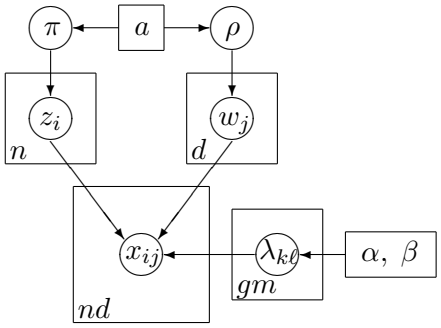

FIGURE 2 – Graphe bayésien du modèle.

L'estimation des labels en ligne et en colonne s'effectue enfin selon la règle du Maximum A Posteriori (MAP).

3 Sélection de modèles

3.1 Critères ICL et BIC

Le travail précédent a été effectué pour un nombre (g, m) de classes fixé, mais choisir un nombre de classes en ligne et colonne pertinent reste un défi d'autant plus que nous nous trouvons dans le cas de données volumineuses. Pour ce faire, un compromis entre la fidélité du modèle aux données et sa complexité doit être fait. Le critère de sélection de modèles ICL (Integrated

Completed Likelihood) pour les données de contingence peut être calculé de manière exacte. Ce type de critère est particulièrement pertinent dans un dessein de classification car contrairement au critère BIC ou AIC, il favorise des composantes bien séparées.

Plus précisément, le critère $ICL(g,m)$ choisit le modèle qui maximise la logvraisemblance complète intégrée $\log p(x,z,w)$. Dans notre cas, il peut être calculé de manière exacte :

$$\begin{aligned} ICL(g,m) &= \log\Gamma(g\times a)+\log\Gamma(m\times a)-(m+g)\log\Gamma(a)+gm\left(\alpha\log\beta-\log\Gamma(\alpha)\right)\\ &- \log\Gamma(n+g\times a)-\log\Gamma(d+m\times a)+\sum_{i,j}\left[-\log x_{ij}!+x_{ij}\log\mu_i+x_{ij}\log\nu_j\right]\\ &+ \sum_{k=1}^{g}\log\Gamma(z_{\cdot k}+a)+\sum_{\ell=1}^{m}\log\Gamma(w_{\cdot\ell}+a)+\sum_{k,\ell}\log\Gamma\left(\alpha+\sum_{i,j}z_{ik}w_{j\ell}x_{ij}\right)\\ &+ \sum_{k,\ell}\left[-\left(\alpha+\sum_{i,j}z_{ik}w_{j\ell}x_{ij}\right)\log\left(\beta+\sum_{i,j}z_{ik}w_{j\ell}\mu_i\nu_j\right)\right]. \end{aligned}$$

Notons qu'en pratique, les variables latentes (z,w) sont remplacées par $(\widehat{z},\widehat{w})$ qui sont les estimateurs fournis par la règle du MAP décrite précédemment.

3.2 Une procédure de parcours du nombre de classes

Étant donné que nous devons parcourir un nombre de classes en ligne et aussi en colonne, le nombre de couples à explorer devient explosif comparé au cas d'un simple mélange classique. Par conséquent, il est nécessaire d'adopter une stratégie d'exploration plus élaborée que celle du parcours exhaustif de chaque couple. Pour ce faire, nous proposons d'adapter l'approche proposée dans le cas d'un mélange simple, par Baudry et Celeux (2015) et basée sur des initialisations récursives.

Supposons que nous devions choisir un modèle de mélange avec un nombre de composantes en ligne et en colonne appartenant respectivement à $\{G_{min},\ldots,G_{max}\}$ et $\{M_{min},\ldots,M_{max}\}$. L'initialisation récursive consiste à découper aléatoirement une des g ou des m composantes en deux afin d'obtenir le modèle $\mathcal{M}_{g+1,m}$ à $(g+1,m)$ composantes et le modèle $\mathcal{M}_{g,m+1}$, puis choisir la solution qui sera la meilleure au sens du critère à maximiser :

- la solution (g_{min},m_{min}) est obtenue minutieusement avec par exemple une procédure telle que l'échantillonneur de Gibbs couplé avec l'algorithme V-Bayes (Keribin *et al.*, 2014).
- Partant de $(g,m)=(g_{min},m_{min})$, la position initiale du modèle $\mathcal{M}_{(g+1,m)}$ et celle du modèle $\mathcal{M}_{(g,m+\ 1)}$ est obtenue en découpant l'une des composantes du modèle $\mathcal{M}_{(g,m)}$ en deux et ce découpage est effectué de manière exhaustive sur chacune des composantes en ligne et en colonne. Enfin, la solution retenue entre le modèle $\mathcal{M}_{(g+1,m)}$ et le modèle $\mathcal{M}_{(g,m+1)}$ sera celle qui maximise le critère de l'algorithme V-Bayes et nous ne retenons que celle-ci, l'autre solution étant écartée (voir Figure 3).

Il s'agit là d'un algorithme glouton.

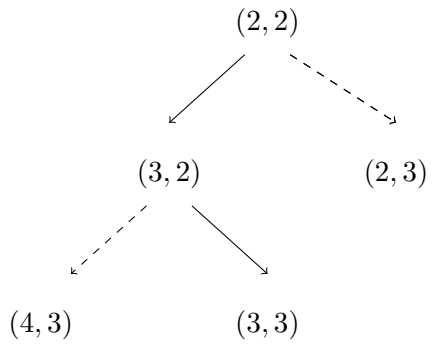

FIGURE 3 – Représentation schématique de l'algorithme de parcours du nombre de classes.

4 Conclusion

Nous terminerons par des illustrations comparant le critère ICL exact présenté précédemment et le critère BIC qui s'écrit dans notre cadre, de la manière suivante :

$$BIC(g,m) \quad = \quad \log p(x;\widehat{\theta}) - \frac{g-1+gm}{2}\log n - \frac{m-1+gm}{2}\log d.$$

Nous ferons également des expérimentations sur données simulées et réelles qui permettront de fournir des recommandations afin que les pharmacologues puissent tirer parti des blocs obtenus. En perspective, une étude plus précise sur chacun des blocs les plus contrastés via les données individuelles et le modèle proposé par Robert *et al.* (2015) pourra être menée.

Bibliographie

[1]Ahmed, I. ; Haramburu, F. ; Fourrier-Réglat, A. ; Thiessard, F. ; Kreft-Jais, C. ; Miremont-Salamé, G. et Tubert-Bitter, P. (2009), Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Statistics in medicine*, 28 (13), 1774–1792.
[2] Baudry, A. et Celeux, G. (2015), EM for mixtures. *Statistics and computing*, 25(4), 713–726.
[3] Govaert, G. et Nadif, M. (2013), Co-clustering. *John Wiley & Sons.*
[4] Govaert, G. et Nadif, M. (2010), Latent block model for contingency table. *Communications in statistics*, 39, 416–425.
[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.
[6] Robert, V. ; Celeux, G. et Keribin, C. (2015), Un modèle statistique pour la pharmacovigilance. *47èmes Journées de Statistique de la SFdS.*