CO-CLUSTERING OF EVOLVING COUNT MATRICES IN PHARMACOVIGILANCE WITH THE DYNAMIC LATENT BLOCK MODEL

Giulia Marchello
Université Côte d'Azur, Inria, CNRS,
Laboratoire J.A.Dieudonné, Maasai team,
Nice, France
giulia.marchello@univ-cotedazur.fr

Audrey Fresse
Université Côte d'Azur,
Department of Clinical Pharmacology,
Pasteur Hospital,
Nice, France
fresse.a@chu-nice.fr

Marco Corneli
Université Côte d'Azur, Inria,
Maison de la Modélisation des Simulations et des Interactions (MSI), Maasai team,
Nice, France
marco.corneli@univ-cotedazur.fr

Charles Bouveyron
Université Côte d'Azur, Inria, CNRS,
Laboratoire J.A.Dieudonné, Maasai team,
Nice, France
charles.bouveyron@univ-cotedazur.fr

ABSTRACT

Pharmacovigilance is a central medical discipline aiming at monitoring and detecting public health events caused by medicines and vaccines. The purpose of this work is to analyze the notifications of adverse drug reactions (ADRs) gathered by the Regional Center of Pharmacovigilance of Nice (France) between 2010 to 2020. As the current expert detection of safety signals is unfortunately incomplete due to the workload it represents, we investigate here an automatized method of safety signal detection from ADRs data. To this end, we introduce a generative co-clustering model, named dynamic latent block model (dLBM), which extends the classical binary latent block model to the case of dynamic count data. The continuous time is handled by partitioning the considered time period, allowing the detection of temporal breaks in the signals. A SEM-Gibbs algorithm is proposed for inference and the ICL criterion is used for model selection. The application to a large-scale ADRs dataset pointed out that dLBM was not only able to identify clusters that are coherent with retrospective knowledge, in particular for major drug-related crises, but also to detect atypical behaviors, which the health professionals were unaware. Thus, dLBM demonstrated its potential as a routine tool in pharmacovigilance.

1 INTRODUCTION

One of the missions of the Regional Centers of Pharmacovigilance (RCPVs) is safety signal detection. However, the method currently used, i.e. manual expert detection of safety signals by the RCPV, despite being unavoidable, has the disadvantage of being incomplete due to its workload. This is why, developing automatized method of safety signal detection is currently a major issue in pharmacovigilance. In such a context, clustering may play an important role in summarizing the information carried out by pharmacovigilance data and identifying patterns of interest. It would be indeed of interest to both cluster the drugs and the adverse reactions to help medical experts in their tasks.

2 THE DYNAMIC LATENT BLOCK MODEL

The main goal of this model is the simultaneous clustering of rows and columns of high-dimensional sparse matrices in a dynamic time framework. The data we consider are organized such that the rows (drugs in pharmacovigilance application) are indexed by $i = 1, \dots N$ and the columns (adversarial effects) by $j = 1, \dots, P$. Moreover, we consider a fixed time period $[0, T]$ during which the total number of rows, N, and columns, P, is fixed. We indicate as $\mathcal{X}(t)$ the $N \times P$ matrix that represents the cumulative number of interactions between i and j at time $t \in [0, T]$. According to the latent block model (Govaert & Nadif, 2010), rows and columns of $\mathcal{X}(t)$ are assumed to be clustered respectively into K and L groups, such that the data belonging to the same block are independent and identically distributed. More formally, the latent structure of $\mathcal{X}(t)$ is identified by:

- $Z := \{z_{ik}\}_{i\in 1,\dots,N, k\in 1,\dots,K}$ represents the clustering of rows into K groups: $\mathcal{A}_1, ..., \mathcal{A}_K$. The row i belongs to cluster $\mathcal{A}_k$ iff $z_{ik} = 1$;
- $W := \{w_{j\ell}\}_{j\in 1,\dots,P, \ell\in 1,\dots,L}$ represents the clustering of columns into L groups: $\mathcal{B}_1, ..., \mathcal{B}_L$. The column j belongs to cluster B_ℓ iff $w_{j\ell} = 1$.

Moreover, Z and W are assumed to be independent and distributed according to multinomial distributions:

$$p(Z|\gamma) = \prod_{k=1}^{K} \gamma_k^{|\mathcal{A}_k|}, \qquad p(W|\rho) = \prod_{\ell=1}^{L} \rho_\ell^{|\mathcal{B}_\ell|},$$

where $\gamma_k = \mathbb{P}\{z_{ik} = 1\}$, $\rho_\ell = \mathbb{P}\{w_{j\ell} = 1\}$, $\sum_{k=1}^{K} \gamma_k = 1$, $\sum_{\ell=1}^{L} \rho_\ell = 1$, and $|\mathcal{A}_k|$ and $|\mathcal{B}_\ell|$ respectively represent the number of rows in cluster $\mathcal{A}_k$ and the number of columns in cluster $\mathcal{B}_\ell$.

Modeling the dynamic framework A possible approach for the dynamic modeling relies on non-homogeneous Poisson processes (NHPPs), thus assuming that $\{\mathcal{X}_{ij}(.)\}_{i,j}$ are independent point processes, with instantaneous intensity functions λ. We further assume that the intensity function only depends on the respective clusters of row i and column j:

$$\mathcal{X}_{ij}(t) \mid z_{ik}, w_{j\ell} = 1 \sim \mathcal{P}\left(\int_0^t \lambda_{k\ell}(u)du\right). \tag{1}$$

In order to ease the understanding of the dynamic model and to make the inference tractable, we also operate a clustering over the time dimension. Let us first introduce a discretization of the considered time interval $[0, T]$, as follows:

$$0 = t_0 < t_1 < \cdots < t_U = T, \tag{2}$$

where the U intervals, $I_u = [t_{u-1}, t_u[$, will also be clustered. The number of interactions between i and j on the time interval I_u can be therefore summarized by:

$$X_{iju} := \mathcal{X}_{ij}(t_u) - \mathcal{X}_{ij}(t_{u-1}), \qquad \forall (i, j, u), \tag{3}$$

where $\mathcal{X}_{ij}(t_u)$ represents the cumulative number of interactions at time t_u between i and j. Since our goal is to perform clustering over the time dimension as well, each time interval $I_1, \dots, I_U$ is also assumed to be assigned to a hidden time cluster $\mathcal{D}_1, \dots, \mathcal{D}_c$. To model the membership to time

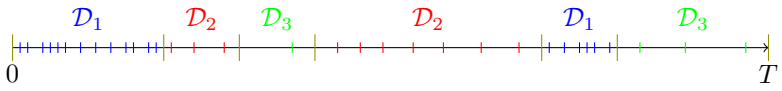

Figure 1: Time clusters.

clusters, a new latent variable S has to be introduced, such that $s_u = c$ if the time interval I_u belongs to the time cluster $\mathcal{D}_c$. As it is shown in Figure 1, it is worth noticing that a specific time cluster can occur more than once in the temporal line when a similar interactivity pattern is repeated in time. Furthermore, as for Z and W, we assume that S follows a multinomial distribution:

$$p(S \mid \delta) = \prod_{c=1}^{C} \delta_c^{|\mathcal{D}_c|}, \tag{4}$$

where $\delta_c = \mathbb{P}\{s_{uc} = 1\}$; $\sum_{c=1}^{C} \delta_c = 1$ and $\mid \mathcal{D}_c \mid$ represents the number of time intervals in the cluster $\mathcal{D}_c$. Once these additional assumptions have been made, we can write:

$$X_{iju} | z_{ik} w_{j\ell} s_{uc} = 1 \sim \mathcal{P}(\lambda_{k\ell c} \Delta_u), \tag{5}$$

where Δ_u indicates the length of the interval I_u.We assume that Δ_u is constant, $\Delta_u = \Delta$. We can finally set $\Delta = 1$ without loss of generality. Thus, it holds that:

$$p(X_{iju} \mid z_{ik} w_{j\ell} s_{uc} = 1, \lambda_{k\ell c}) = \left(\frac{(\lambda_{k\ell c})^{X_{iju}}}{X_{iju}!} \exp\left(-\lambda_{k\ell c}\right) \right). \tag{6}$$

It is now possible to write the complete data likelihood of the model:

$$p(X, Z, W, S | \gamma, \rho, \delta, \lambda) = p(Z|\gamma) p(W|\rho) p(S|\delta) p(X|Z, W, S, \lambda), \tag{7}$$

where $p(Z|\gamma)$, $p(W|\rho)$ and $p(S|\delta)$ were defined in the previous section. The conditional distribution of X, given Z, W, and S, can be easily obtained from Eq. equation 6 by independence:

$$p(X|Z, W, S, \lambda) = \prod_{k,\ell,c} \left(\frac{(\lambda_{k\ell c})^{R_{k\ell c}}}{P_{k\ell c}} \exp\left(- \mid \mathcal{A}_k \mid\mid \mathcal{B}_\ell \mid\mid \mathcal{D}_c \mid \lambda_{k\ell c}\right) \right), \tag{8}$$

where $R_{k\ell c} = \sum_{i=1}^{N} \sum_{j=1}^{P} \sum_{u=1}^{U} z_{ik} w_{j\ell} s_{uc} X_{iju}$ and $P_{k\ell c} = \prod_{i=1}^{N} \prod_{j=1}^{P} \prod_{u=1}^{U} (z_{ik} w_{j\ell} s_{uc} X_{iju})!$.

Model inference In the co-clustering case, the EM algorithm is computationally unfeasible, to go through this limitation, we propose to approximate it through a Gibbs sampler within the E-step. Such an approach was proposed by Keribin et al. (2010) and exploited, for instance, by Bouveyron et al. (2018) for the functional latent block model (funLBM). Thanks to the Gibbs sampler within the SE step a partition for Z, W and S is generated without computing the joint distribution. The algorithm starts with initial values for the parameter set $\theta^{(0)}$, the column clusters $W^{(0)}$ and the time clusters $S^{(0)}$. Regarding the burn-in period, after a certain number of iterations of the algorithm, we can obtain the final parameters estimation by computing the mean of the sampled distribution. The optimal values for Z, W and S are estimated by the mode of their sample distributions.

Model selection Up to now, we have assumed that the number of row clusters (K), column clusters (L) and time clusters (C) was known. However, for real data sets, this assumption is of course unrealistic. For this reason, our purpose in this section is to define a model selection criterion that can automatically identify the optimal number of clusters. We rely on ICL (Integrated Completed Likelihood, Biernacki et al. (2000)):

$$\begin{aligned} ICL(K, L, C) = \log p(X, \hat{Z}, \hat{W}, \hat{S}; \hat{\theta}) - \frac{K-1}{2} \log N + \\ - \frac{L-1}{2} \log P - \frac{C-1}{2} \log U - \frac{KLC}{2} \log(NPU) \end{aligned} \tag{9}$$

The triplet $(\hat{K}, \hat{L}, \hat{C})$ that leads to the highest value for the ICL is considered as the most meaningful for those data.

3 ANALYSIS OF THE ADVERSE DRUG REACTION DATASET

This section considers a large dataset consisting of ADR data collected by the Regional Center of Pharmacovigilance (RCPV), located in the University Hospital of Nice (France). The center covers an area of over 2.3 million inhabitants. A time horizon of 10 years is considered, from January 1^{st}, 2010 to September 30^{th}, 2020, the unity measure for time intervals is a month ($\Delta_u = \Delta =$ 1 month). The overall dataset is made of by 44,269 declarations. We only considered drugs and ADRs that were notified more than 10 times over the 10 years. During this period, an extremely uncommon behavior happened in the progress of notifications to the RCPV. In fact, in 2017 an unexpected rise of reports for ADRs happened concerning a specific drug called Lévothyrox®. This has been marketed in France for about 40 years as a treatment for hypothyroidism and, in 2017, a new formula was introduced on the market. The Lévothyrox® case had an extremely high media coverage in France: Lévothyrox® spontaneous reports represent almost the 90% of all the spontaneous notifications that the RCPV received in 2017 Viard et al. (2019). Behind those very visible effects, many ADR signals need to be detected for obvious public health reasons. In particular, those data also contain ADR reports regarding Médiator®, which is here far less visible, but also led to many avoidable serious cardiovascular diseases. This is why, we expect dLBM to be a useful tool to reveal such hidden signals.

3.1 SUMMARY OF THE RESULTS

We have run dLBM for different values of K, L and C, we tested row (here drugs), column (here ADRs) and time groups ranging from 1 to 12. The ICL criterion identified the optimal values as: $\hat{K} = 7$, $\hat{L} = 10$, $\hat{C} = 6$. Figure 2 shows the frequency of the declarations received by the RCPV from 2010 to 2020, sorted by month, where the colors represent the identified time clusters. Figure 3 shows the evolution of the relationship between drug clusters and ADR clusters over time. In fact, each panel represents a cluster of drugs and within them each line identifies a cluster of ADRs and its intensity changes over time. In this application to pharmacovigilance, dLBM proved to be a very useful tool for identifying phenomena that would have been difficult to detect otherwise, even by an expert eye. In fact, dLBM revealed that in addition to Lévothyrox® health crisis, which was the one with the widest media coverage, two other major events have occurred. The first one concerning Médiator®, which took place in 2009-2010, and the second one concerning Mirena®, which took place in the first half of 2017. From a more in-depth analysis of the time clusters, one can easily notice on Figure 2 that the segmentation proposed by the algorithm confirms our knowledge about the previous mentioned health scandals while revealing a time structure more complex than expected. In fact, while cluster 1 and cluster 2 include various time intervals, cluster 3 clearly refers to the health crisis due to the Mirena® scandal while cluster 4 relates to the peak period in the Lévothyox® crisis. Time clusters 5 and 6 refer to the final stage of the Lévothyrox® crisis, when generics were introduced to the market. It is worth noticing that without the dLBM application it would have been impossible to detect the presence of other health scandal just before the one of Lévothyrox®. In fact, looking at Figure 2, one can see that the increase of declarations during the Mirena® health crisis are completely masked by the Lévothyrox® ones. The clusters of drugs identified by the algorithm are also coherent with retrospective knowledge and adequately represent the variety of drugs present in the dataset. In particular, cluster 1, cluster 6 and cluster 7 are very specific, with one element only: they correspond respectively to lévothyroxine (Lévothyrox® and generics), benfluorex (Médiator®) and lévonorgestrel (Mirena®). Moreover, cluster 2 contains the five most frequently reported drugs and cluster 5 contains other common drugs, while cluster 4 is very large and heterogeneous, with drugs that are rarely reported and finally cluster 3 contains drugs that cause bleeding. Concerning the clusters of ADRs, cluster 3 and cluster 8 contain the most frequently notified ADRs. Cluster 1 contains recurring ADRs but less than the other two previously mentioned. Cluster 2 and cluster 4 respectively include the most and the less frequent bleeding related ADRs. Cluster 7 is composed of ADRs clearly related to Lévothyrox® and Mirena® (e.g hair loss, cramps, insomnia, etc.). In cluster 10 there are general ADRs, although it contains some ADRs specifically related to Lévothyrox® and Médiator®. Finally, cluster 5, 6 and 9 contain more general and nonspecific ADRs. In addition, dLBM was also able to put in light some unexpected variations of notifications such as an under-notification of bleeding related ADRs during Lévothyrox® crisis. Another thing that dLBM has highlighted is the existence of 3 different phases during the Lévothyrox® crisis corresponding to the reporting peak, the marketing period of generics and the end of the crisis, respectively. Those

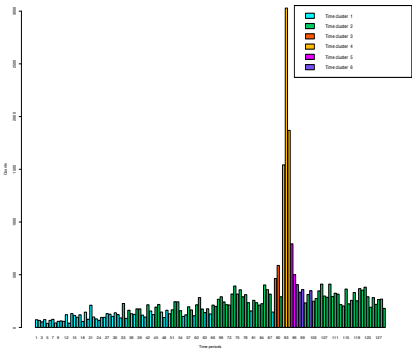

Figure 2: Reports received by the RCPV, colors represent the time clusters.

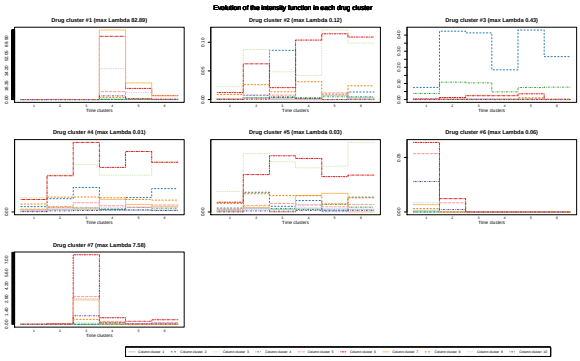

Figure 3: Evolution of the relation between each drug cluster and the all ADR clusters over time. Each color corresponds to a different ADR cluster.

phases were not noticed by the RCPV staff during the Lévothyrox® crisis. In general, we can conclude that dLBM could be extremly useful as a routine tool for signal detection, since it might help health professionals to identify structural changes or patterns of interest and, perhaps, prevent some of the consequences a health crisis can lead to.

REFERENCES

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

Charles Bouveyron, Laurent Bozzi, Julien Jacques, and François-Xavier Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):897–915, 2018.

Gérard Govaert and Mohamed Nadif. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425, 2010.

Christine Keribin, Gérard Govaert, and Gilles Celeux. Estimation d'un modèle à blocs latents par l'algorithme sem. 2010.

Delphine Viard, Nadège Parassol-Girard, Serena Romani, Elise Van Obberghen, Fanny Rocher, Sofia Berriri, and Milou-Daniel Drici. Spontaneous adverse event notifications by patients subsequent to the marketing of a new formulation of levothyrox® amidst a drug media crisis: atypical profile as compared with other drugs. *Fundamental & clinical pharmacology*, 33(4):463–470, 2019.