Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance

Valérie Robert

NNT : 2017SACLS111

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS
Laboratoire B2PHI, UMR 1181 INSERM PASTEUR UVSQ

Spécialité de doctorat : Mathématiques appliquées

Valérie ROBERT

Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance.

Date de soutenance : 06 Juin 2017

Après avis des rapporteurs : CHANTAL GUIHENNEUC-JOUYAUX (Université Paris Descartes)
JULIEN JACQUES (Université Lyon 2)

Jury de soutenance :

CHARLES BOUVEYRON	Université Paris Descartes	Président du jury
GILLES CELEUX	INRIA Saclay	Directeur de thèse
CHRISTOPHE GIRAUD	Université Paris Sud	Examinateur
CHANTAL GUIHENNEUC-JOUYAUX	Université Paris Descartes	Rapporteure
JULIEN JACQUES	Université Lyon 2	Rapporteur
CHRISTINE KERIBIN	Université Paris Sud	Co-directrice de thèse
PASCALE TUBERT-BITTER	INSERM, PASTEUR, UVSQ	Invitée

Thèse préparée au
Laboratoire de Mathématiques d'Orsay
Bât. 425 Université Paris Sud
91405 Orsay CEDEX

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance.

Mots Clefs : pharmacovigilance – classification croisée – modèles de mélange – algorithmes bayésiens – approximation variationnelle – sélection de modèles

Résumé :
Cette thèse regroupe des contributions méthodologiques à l'analyse statistique des bases de données de pharmacovigilance. Les difficultés de modélisation de ces données résident dans le fait qu'elles produisent des matrices souvent creuses et de grandes dimensions. La première partie des travaux de cette thèse porte sur la classification croisée du tableau de contingence de pharmacovigilance à l'aide du modèle des blocs latents de Poisson normalisé. L'objectif de la classification est d'une part de fournir aux pharmacologues des zones intéressantes plus réduites à explorer de manière plus précise, et d'autre part de constituer une information a priori utilisable lors de l'analyse des données individuelles de pharmacovigilance. Dans ce cadre, nous détaillons une procédure d'estimation partiellement bayésienne des paramètres du modèle et des critères de sélection de modèles afin de choisir le modèle le plus adapté aux données étudiées. Les données étant de grandes dimensions, nous proposons également une procédure pour explorer de manière non exhaustive mais pertinente, l'espace des modèles en coclustering. Enfin, pour mesurer la performance des algorithmes, nous développons un indice de classification croisée calculable en pratique pour un nombre de classes élevé. Les développements de ces outils statistiques ne sont pas spécifiques à la pharmacovigilance et peuvent être utile à toute analyse en classification croisée. La seconde partie des travaux de cette thèse porte sur l'analyse statistique des données individuelles, plus nombreuses mais également plus riches en information. L'objectif est d'établir des classes d'individus selon leur profil médicamenteux et des sous-groupes d'effets et de médicaments possiblement en interaction, palliant ainsi le phénomène de coprescription et de masquage que peuvent présenter les méthodes existantes sur le tableau de contingence. De plus, l'interaction entre plusieurs effets indésirables y est pris en compte. Nous proposons alors le modèle des blocs latents multiple qui fournit une classification croisée simultanée des lignes et des colonnes de deux tableaux de données binaires en leur imposant le même classement en ligne. Nous discutons des hypothèses inhérentes à ce nouveau modèle et nous énonçons des conditions suffisantes de son identifiabilité. Ensuite, nous présentons une procédure d'estimation de ses paramètres et développons des critères de sélection de modèles associés. De plus, un modèle de simulation numérique des données individuelles de pharmacovigilance est proposé et permet de confronter les méthodes entre elles et d'étudier leurs limites. Enfin, la méthodologie proposée pour traiter les données individuelles de pharmacovigilance est explicitée et appliquée à un échantillon de la base française de pharmacovigilance entre 2002 et 2010.

Title : Coclustering for the analysis of pharmacovigilance large datasets

Keys words : Pharmacovigilance – Coclustering – Mixture Models – Bayesian Algorithms – Variational Approximation – Model Selection

Abstract :
This thesis gathers methodological contributions to the statistical analysis of large datasets in pharmacovigilance. The pharmacovigilance datasets produce sparse and large matrices and these two characteritics are the main statistical challenges for modelling them. The first part of the thesis is dedicated to the coclustering of the pharmacovigilance contingency table thanks to the normalized Poisson latent block model. The objective is on the one hand, to provide pharmacologists with some interesting and reduced areas to explore more precisely. On the other hand, this coclustering remains a useful background information for dealing with individual database. Within this framework, a parameter estimation procedure for this model is detailed and objective model selection criteria are developed to choose the best fit model. Datasets are so large that we propose a procedure to explore the model space in coclustering, in a non exhaustive way but in a relevant one. Additionnally, to assess the performances of the methods, a convenient coclustering index is developed to compare partitions with high numbers of clusters. The development of these statistical tools are not specific to pharmacovigilance and can be used for any coclustering issue. The second part of the thesis is devoted to the statistical analysis of the large individual data, which are more numerous but also provides even more valuable information. The aim is to produce individual clusters according their drug profiles and subgroups of drugs and adverse effects with possible links, which overcomes the coprescription and masking phenomenons, common contingency table issues in pharmacovigilance. Moreover, the interaction between several adverse effects is taken into account. For this purpose, we propose a new model, the Multiple Latent Block Model (MLBM) which enables to cocluster two binary tables by imposing the same row ranking. Assertions inherent to the model are discussed and sufficient identifiability conditions for the model are presented. Then a parameter estimation algorithm is studied and objective model selection criteria are developed. Moreover, a numeric simulation model of the individual data is proposed to compare existing methods and study its limits. Finally, the proposed methodology to deal with individual pharmacovigilance data is presented and applied to a sample of the French pharmacovigilance database between 2002 and 2010.

« *N'abandonne jamais.* »
J. Dawson

Table des matières

Notations

Partie I

Indices

- H : nombre de classes en ligne pour le tableau c ,
- h : indice des classes en colonne de c allant de 1 à H,
- L: nombre de classes en colonne pour le tableau c,
- ℓ : indice des classes en colonne de c allant de 1 à L,
- J : nombre de lignes du tableau c,
- j : indice des lignes du tableau c, allant de 1 à J,
- K : nombre de colonnes du tableau c,
- k : indice des colonnes du tableau c, allant de 1 à K.

Variables, paramètres

- c tableau de contingence médicaments/effets de taille $J \times K$,
- $\gamma_{h\ell}$ paramètre de la loi du bloc (h, ℓ) de c,
- μ_j paramètre de normalisation modélisant l'effet ligne,
- ν_k paramètre de normalisation modélisant l'effet colonne,

- ρ_h probabilité pour une ligne de c d'appartenir à la classe h,
- τ_ℓ probabilité pour une colonne de c d'appartenir à la classe ℓ,
- $\theta = (\gamma, \pi, \rho, \tau)$ paramètre du modèle à estimer,
- (v, w) matrice des partitions en ligne et en colonne pour c,
- v_{jh} case de la matrice v indiquant l'appartenance de la ligne j de c à la classe h,
- $w_{k\ell}$ case de la matrice w indiquant l'appartenance de la colonne k de c à la classe ℓ.

Partie II

Indices

- G : nombre de classes en ligne,
- g : indice des classes en ligne allant de 1 à G,
- H : nombre de classes en colonne pour le premier tableau x,
- h : indice des classes en colonne de x allant de 1 à H,
- L: nombre de classes en colonne pour le deuxième tableau y,
- ℓ : indice des classes en colonne de y allant de 1 à L,
- n : nombre de lignes ou d'individus,
- i : indice des lignes allant de 1 à n,
- J : nombre de colonnes du premier tableau x,
- j : indice des colonnes du premier tableau x, allant de 1 à J,
- K : nombre de colonnes du deuxième tableau y,

- k : indice des colonnes du deuxième tableau y, allant de 1 à K.

Variables, paramètres

- x tableau des individus et des médicaments de taille $n \times J$,
- y tableau des individus et des effets de taille $n \times K$,
- α_{gh} paramètre de la loi du bloc (g, h) de x,
- $\beta_{g\ell}$ paramètre de la loi du bloc (g, ℓ) de y,
- π_g probabilité pour une ligne d'appartenir à la classe g,
- ρ_h probabilité pour une colonne de x d'appartenir à la classe h,
- τ_ℓ probabilité pour une colonne de y d'appartenir à la classe ℓ,
- $\theta = (\alpha, \beta, \pi, \rho, \tau)$ paramètre du modèle à estimer,
- (z, v, w) matrice des partitions en ligne et en colonne pour x et y,
- z_{ig} case de la matrice z indiquant l'appartenance de la ligne i à la classe g,
- v_{jh} case de la matrice v indiquant l'appartenance de la colonne j de x à la classe h,
- $w_{k\ell}$ case de la matrice w indiquant l'appartenance de la colonne k de y à la classe ℓ.

Acronymes

- LBM : Latent Block Model,
- MLBM : Multiple Latent Block Model,
- ICL : Integrated Complete Likelihood,
- BIC : Bayesian Information Criterion.

- MAP : Maximum A Posteriori.
- VEM : Variational Expectation Maximisation.

1

Introduction

1.1 Contexte et problématique

La pharmacovigilance est la surveillance des médicaments et la prévention du risque d'effet indésirable résultant de leur utilisation et s'opère en deux phases. Durant la première phrase, des essais cliniques sont effectués avant l'autorisation de mise sur le marché des médicaments. Mais ces études sont généralement conduites sur un panel homogène de taille fixée et ainsi pour certaines catégories de personnes (femmes enceintes, diabétiques ...) peu d'informations sont disponibles. De plus, ces observations sont effectuées sur une durée limitée et les effets indésirables qui surviennent après une longue période de latence sont ignorés. Enfin, les conditions d'utilisation des médicaments sont souvent différentes de celles étudiées dans les cadres expérimentaux.

Ainsi, dans beaucoup de pays, s'opère une deuxième phase de pharmacovigilance qui vise à détecter le plus tôt possible les effets indésirables non répertoriés et induits par les médicaments après leur mise sur le marché.

Dans ce cadre, le système de pharmacovigilance repose sur la déclaration par des professionnels de santé de la survenue d'événements indésirables dont la cause suspectée est médicamenteuse. L'accumulation des ces notifications spontanées permet alors aux pharmacologues d'exhiber des effets indésirables médicamenteux.

Malheureusement, la base de données ainsi générée présente des biais étant donné que tous les effets indésirables ne sont pas rapportés aux instances de pharmacovigilance. Cette sous-notification peut même être importante dans le cas d'effets indésirables graves (Almenoff et al. (2007),van der Heijden et al. (2002)).

Cependant, l'analyse de ces notifications reste le moyen le plus rapide d'identifier des associations potentielles entre effets indésirables et médicaments et constitue une aide précieuse pour les instances de pharmacovigilance dans leurs prises de décisions.

1.1.1 Le système de pharmacovigilance

Le système de pharmacovigilance français a été mis en place en 1979 et repose actuellement sur 31 centres régionaux de pharmacovigilance (CRPV, voir figure 1.1). Ils sont en charge de l'enregistrement des notifications au sein de la base nationale coordonnée par l'unité de pharmacovigilance de l'Agence Française de Sécurité Sanitaire des Produits de Santé (Afssaps).

Figure 1.1 – Les points rouges représentent les centres régionaux de pharmacovigilance français.

C'est aussi dans ces centres que s'exerce une partie de la surveillance ; les cas suspectés étant par la suite discutés par le comité technique de pharmacovigilance lors de réunions mensuelles au siège de l'Afssaps. Entre 2002 et 2010, les données de pharmacovigilance françaises atteignent le nombre d'environ 200 000 notifications.

Par ailleurs, il existe d'autres bases plus conséquentes que la base française : la base américaine coordonnée par la Food and Drug Administration (FDA) et la base de l'Organisation Mondiale de la Santé administrée par l'Uppsala Monitoring Center en Suède, qui, en décembre 2004, contenaient respectivement environ 2.6 et 3.7 millions de notifications (Almenoff et al. (2007)).

1.1.2 Structure des données analysées

1.1.2.a Données individuelles

Les données de pharmacovigilance se présentent sous la forme de deux matrices binaires très creuses notées $x = (x_{ij})$ et $y = (y_{ik})$ de tailles respectives $n \times J$ et $n \times K$ (voir figure 1.2) telles que :

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

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

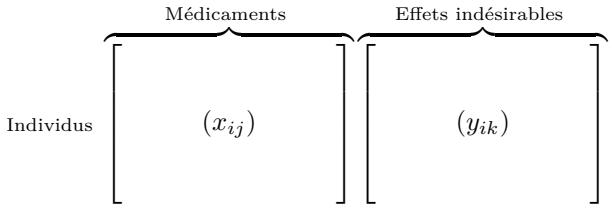


Figure 1.2 – Représentation des données individuelles de pharmacovigilance

Remarque. Les matrices x et y ne possèdent aucune ligne non nulle, et tous les médicaments et effets indésirables présents dans la base on été notifiés au moins une fois.

1.1.2.b Tableau de contingence

À partir de ces données individuelles, nous pouvons effectuer une projection de celles-ci en un tableau de contingence noté $c = (c_{jk})$ de taille $J \times K$ croisant l'ensemble des médicaments et les événements indésirables impliqués au moins une fois dans une notification spontanée (voir figure 1.3). Ce tableau est aussi caractérisé par une grande proportion de cellules vides et est obtenu à partir des données individuelles par la formule suivante :

$$c = {}^t x \times y,$$

où t désigne l'opération transposée.

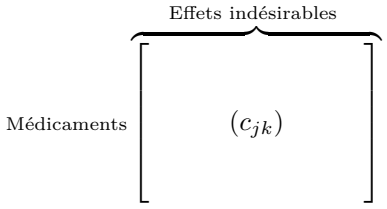


Figure 1.3 – Représentation des données de pharmacovigilance en tableau de contingence.

1.1.2.c Codage des événements indésirables et des médicaments

La taille des matrices des données individuelles et du tableau de contingence dépend fortement du degré de précision utilisé pour le codage des événements indésirables et des médicaments. En effet, les médicaments sont codés dans la base de pharmacovigilance selon la classification ATC (Anatomical Therapeutical Clinical, Miller and Britt (1995)) et les effets

indésirables sont codés selon la classification MedDRA (Medical Dictionnary for Regulatory Activities, Brown et al. (1999)). La classification ATC est arborescente et présente cinq niveaux de précision allant du groupe anatomique (code à une lettre notée ATC1, voir tableau 1.1) à la dénomination commune internationale (code à sept caractères noté ATC7). Les données de pharmacovigilance sont généralement codés au premier niveau de regroupement de la classification ATC, soit ATC5.

A	Système digestif et métabolisme
B	Sang et organes hématopoiétiques
C	Système cardio-vasculaire
D	Dermatologie
G	Système génito-urinaire et hormones sexuelles
H	Hormones systémiques, à l'exclusion des hormones sexuelles et des insulines
J	Anti-infectieux (usage systémique)
L	Antinéoplasiques et agents immunomodulants
M	Système musculo-squelettique
N	Système nerveux
P	Antiparasitaires, insecticides et répulsifs
R	Système respiratoire
S	Organes sensoriels
V	Divers

Table 1.1 – Liste des 14 groupes principaux de la classification ATC.

La classification MedDRA qui concerne les effets indésirables présente également cinq niveaux de précision :

- System Organ Class (SOC),
- High-Level Group Terms (HLGT),
- High-Level Terms (HLT),
- Preferred Terms (PT),
- Lower-Level Terms (LLT).

Les données de pharmacovigilance sont généralement codées en High-Level Terms (HLT).

1.2 Motivations et objectifs de la thèse

Face à cette émergence de données massives, des méthodes statistiques de génération automatique de signaux ont été développées depuis une vingtaine d'années. Les premières méthodes utilisent plutôt les projections des données individuelles en tableaux de contingence car effectuer une étude globale sur les données individuelles reste un défi majeur au vu de leurs tailles. Notons que cela présuppose d'une homogénéité des individus à l'origine des notifications. Les méthodes de détection automatique les plus utilisées sont les méthodes *Proportional Reporting Ratio* (PRR, Evans et al. (2001)), *Reporting Odds Ratio* (ROR, van Puijenbroek et al. (2002)), *Bayesian Confidence Propagation Neural Network* ($BCPNN$, Bate et al. (1998); Norén et al. (2006)), et *(Multi-item) Gamma Poisson Shrinker* ($(M)GPS$, DuMouchel (1999); DuMouchel and Pregibon (2001)). La méthode PRR est utilisée par le

système de pharmacovigilance anglais et également par le système européen Eudravigilance. La méthode ROR est expérimentée sur la base nationale des Pays-Bas. Ensuite, la méthode $BCPNN$ est exploitée sur la base de l'OMS alors que la méthode $MGPS$ est utilisée par la FDA (Food Drug Administration). Enfin, la France n'a pas encore recours à l'une de ces méthodes décrites précédemment mais le travail de Ahmed (2009), Marbac et al. (2016) témoignent de la volonté de mettre un tel système en place à long terme.

L'objectif de ces méthodes est de construire une mesure de dissimilarité et vise ainsi à détecter les couples effet-médicament dont la présence est anormalement fréquente (vis-à-vis d'un seuil souvent arbitraire) par rapport à ce qui est attendu compte tenu de l'information présente dans le reste de la base (voir figure 1.4). Ces procédures diffèrent sur la mesure de dissimilarité choisie, la modélisation à l'origine de cette mesure ainsi que sur les seuils de génération d'alerte. De plus, les méthodes $BCPNN$ et GPS possèdent un degré de raffinement plus grand et utilisent une modélisation plus complexe des données.

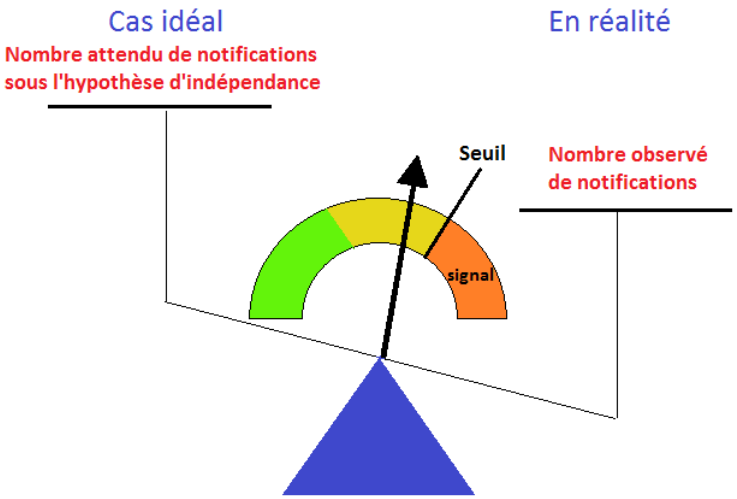


Figure 1.4 – Représentation schématique des méthodes de génération automatique de signaux.

Enfin, plus récemment, des méthodes ont été développées sur les données individuelles et non plus agrégées. Ainsi, Caster et al. (2010) insiste sur le fait que ces projections des données sur des tableaux de contingence montrent quelques faiblesses vis-à-vis des phénomènes de *coprescription* de médicaments et d'*effet de masquage*.

En effet, soient un médicament 1 responsable d'un effet et un médicament 2 non responsable de celui-ci souvent co-prescrits ensemble. Les méthodes travaillant sur le tableau de contingence, peuvent alors attribuer l'effet au médicament 2 car certes, cette coprescription est visible sur les données individuelles, mais cette information est perdue pour le tableau de contingence.

L'effet de masquage, se manifeste quand certaines associations médicaments/effet sont fortement notifiés dans le tableau de contingence. Lors du calcul de la mesure de disproportionnalité de couples nettement notifiés, ces associations interviennent et peuvent alors fausser l'estimation de cette mesure et masquer ainsi des signaux potentiels.

Ainsi, pour pallier ces problèmes, le retour aux données individuelles est préconisé. La probabilité de survenue d'un effet indésirable conditionnellement à la prise de médicaments peut être alors modélisée par une régression logistique parcimonieuse (Caster et al. (2010),

Marbac et al. (2016)) ou par une régression logistique classique avec une étape de présélection de médicaments d'intérêt (Harpaz et al. (2013)). Ces méthodes procèdent effet par effet et ne tiennent pas compte de plusieurs effets indésirables simultanément.

Dans tous les cas, du fait des limites des notifications spontanées énoncées plus haut et de la nature essentiellement exploratoire de ces analyses, les signaux ainsi générés doivent être examinés par les pharmacologues pour en évaluer la pertinence. Il s'agit donc plutôt d'outils complémentaires à la veille opérée par les pharmacologues (Ahmed (2009)).

Ainsi, aucune de ces méthodes n'est définie comme la méthode de référence car, par manque de jeux de données *benchmark* en pharmacovigilance et de consensus sur l'évaluation des méthodes, leur comparaison reste un défi.

Quant aux données réelles, des jeux de référence commencent à émerger tels que celui proposé par Ryan et al. (2013). En effet, le jeu de référence $OMOP$ (Ryan et al. (2013)) contient des témoins positifs et des témoins négatifs pour quatre effets indésirables d'intérêt : lésions hépatiques aigües (ALI), lésions rénales aigües (AKI), infection aigüe du myocarde (AMI), et saignements gastro-intestinaux supérieurs (GIB). Dans cet ensembe de référence, 399 couples témoins sont disponibles et parmi eux, 165 sont des témoins positifs et 234 sont des témoins négatifs. Ryan et al. (2013) précisent que la plupart des témoins positifs pour les effets AKI et GIB sont issus de résultats d'essais cliniques randomisés alors que la majorité des témoins positifs pour AMI et ALI sont basés sur des études de cas publiés. Ainsi, il n'est évidemment pas exhaustif et dans la base française que nous allons étudier, seuls 198 témoins sont présents, c'est-à-dire avec au moins une notification (voir tableau 1.2).

Par ailleurs, des protocoles de simulation du tableau de contingence (Roux et al. (2005)) ont été proposés et d'autres pour simuler des données individuelles mais ne concernent qu'un seul effet indésirable donné (Ahmed et al. (2016)).

Effet indésirable	AMI	GIB	ALI	AKI
Témoin positif	14	17	70	16
Témoin négatif	13	28	17	23

Table 1.2 – Nombre de signaux de l'OMOP présents (avec au moins une notification) dans la base française de pharmacovigilance entre 2002 et 2010.

En conclusion, les méthodes existantes travaillant en projection des données individuelles sur tableaux de contingence possèdent certaines faiblesses (seuils arbitraires, effet de co-prescription, effet de masquage) et présupposent une homogénéité des individus à l'origine des notifications. Quant à l'émergence des procédures sur les données individuelles, aucune d'entre elles ne prend en compte l'interaction de plusieurs effets indésirables entre eux.

Dans cette thèse, nous cherchons à réaliser une modélisation statistique des données individuelles de pharmacovigilance pour permettre de détecter des associations potentielles entre un médicaments et un effet, mais aussi l'association de plusieurs médicaments entraînant un ou plusieurs effets. Il est ainsi important que le modèle prenne en compte la dimension hétérogène du problème et le côté multidimensionnel de la variable effet indésirable, avec la difficulté d'être confrontés à des données massives.

1.3 Démarches et contributions

La démarche consiste dans un premier temps, à analyser les méthodes existantes de traitement des données de pharmacovigilance et à identifier leurs avantages et limitations.

Dans un deuxième temps, nous proposons de tirer parti de l'information du tableau de contingence. Pour ce faire, nous proposons d'effectuer une classification croisée du tableau de contingence et d'un tableau de contingence réduit jusque là peu étudié, en considérant les individus qui ont pris un seul médicament et un seul effet. Celle-ci permet d'obtenir une première information résumée et de fournir des zones intéressantes d'intérêt du tableau pour les pharmacologues.

Puis nous traitons les données individuelles en développant un nouveau modèle statistique, le modèle des blocs latents multiple $MLBM$ 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 leurs profils médicamenteux et des sous-groupes d'effets et de médicaments qui sont peut-être liés. Les résultats obtenus sur le tableau de contingence sont alors utilisés comme information a priori lors du traitement des données individuelles par ce modèle.

Traitement du tableau de contingence

La modélisation est basée sur le modèle des blocs latents (LBM, Govaert and Nadif (2007)). Étant donné un tableau, le principe est d'obtenir une classification v des lignes en H groupes et une classification w des colonnes en L groupes de telle sorte que les blocs obtenus soient homogènes. Conditionnellement à l'appartenance à un bloc, les cellules d'un bloc sont alors le résultat de variables aléatoires indépendantes et identiquement distribuées suivant une même loi inconnue. Ces lois appartiennent à une même famille paramétrique. Le LBM peut alors s'appliquer à plusieurs types de données : binaires (Govaert and Nadif (2007)) en utilisant la loi de Bernoulli, réelles en utilisant la loi gaussienne (Lomet (2012)), catégorielles en utilisant la loi multinomiale (Keribin et al. (2015)), ordinales en utilisant le modèle BOS (*Binary Ordinal Search*, Biernacki and Jacques (2015)), fonctionnelles en utilisant la base fournie par l'ACP fonctionnel (travail en cours de Ben Slimen et al. (2016)) et de comptage (Govaert and Nadif (2013)) en utilisant la loi de Poisson. Nous considérons cette dernière approche pour modéliser le tableau de contingence $c = (c_{jk})_{J\times K}$.La densité conditionnelle pour une observation c_{jk} du bloc $k\ell$ s'écrit alors,

$$\phi(c_{jk};\mu_j\nu_k\gamma_{h\ell}) = e^{-\mu_j\nu_k\gamma_{h\ell}}\frac{(\mu_j\nu_k\gamma_{h\ell})^{c_{jk}}}{c_{jk}!},$$

où μ_j représente l'effet ligne (de la sorte, deux lignes proportionnelles seront mises dans le même bloc), ν_k représente l'effet colonne et $\gamma_{h\ell}$ représente l'interaction à l'intérieur du bloc $h\ell$.

Estimation et discussion autour des hyperparamètres. Comme la vraisemblance du modèle n'est pas calculable analytiquement à cause de cette dépendance entre v et w d'une part, et d'autre part, que l'algorithme EM classique pour estimer ces modèles à variables cachées, est également inutilisable en pratique, nous adaptons la procédure de Keribin et al. (2015) développée dans le cas de données catégorielles : le couplage de l'échantillonneur de Gibbs avec l'algorithme *V-Bayes* qui pallie à la fois le problème d'initialisation et le problème de dégénérescence des classes que peuvent rencontrer certains algorithmes dérivés. Cette approche bayésienne nécessite des lois a priori les moins informatives possibles sur les

paramètres du modèle que nous choisissons conjuguées :

$$\rho \sim \mathcal{D}(a,..,a) \quad \tau \sim \mathcal{D}(a,..,a), \quad \text{et} \quad \gamma_{h\ell}|\alpha,\beta \sim \Gamma(\alpha,\beta),$$

où $\rho = \mathbb{P}(v = 1)$ et $\tau = \mathbb{P}(w = 1)$. Le choix des hyperparamètres est un défi important d'autant plus qu'il n'existe pas d'hyperparamètres α, β fournissant une loi non informative pour les γ. Nous réalisons une étude détaillée de sensibilité des algorithmes en se basant sur des données simulées. Nous avons alors conforté le choix de a=4 qui limite au mieux le problème de dégénérescence des classes en jouant le rôle d'un paramètre de régularisation dans l'algorithme *V-Bayes* (Keribin et al. (2015)) et prendre $\alpha = 1$ a également un effet bénéfique sur ce phénomène. Enfin, afin de réaliser un compromis entre choisir une loi a priori la moins informative possible et les résultats obtenus, nous préconisons de choisir $\beta = 0.01$.

Sélection de modèles : aspects théoriques et algorithmiques. Le nombre adéquat de classes en ligne et en colonne doit être estimé et nous proposons de le sélectionner de manière automatique. Dans cette étude, différents critères de sélection de modèles ont été développés et testés (*Integrated Completed Loglikelihood*, ICL (Biernacki et al. (2000)); *Bayesian Criterion Information*, BIC Schwarz et al. (1978). De plus, nous proposons une nouvelle procédure *Bi-KM1*, basée sur les initialisations récursives (Baudry and Celeux (2015)), qui permet de parcourir de manière non exhaustive le nombre de classes en ligne et en colonne (voir figure 1.5). Ainsi, au lieu de parcourir une grille de taille $H_{max} \times L_{max}$, ce qui est extrêmement coûteux, l'algorithme visite dans le pire des cas $H_{max} + L_{max}$ couples de classes en ligne et en colonne, ce qui est nettement moindre.

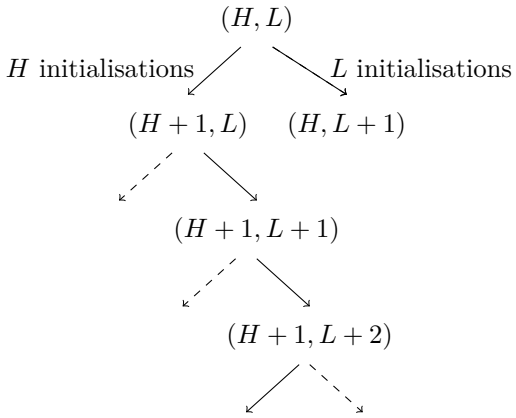


Figure 1.5 – Représentation schématique de l'algorithme *Bi-KM1*.

En se basant sur des données simulées, nous testons ensuite leur capacité à estimer le vrai nombre de classes en ligne et en colonne. De plus, une comparaison détaillée avec un autre algorithme d'optimisation d'ICL (GS, Wyse et al. (2016)) y est proposée. La comparaison s'effectue sur données simulées et réelles. Pour comparer les partitions entre elles, nous proposons un indice de classification croisée inspiré de Hubert and Arabie (1985) et adapté au problème des données massives calculable en pratique pour un grand nombre de classes en ligne et en colonne (contrairement à celui proposé par Lomet (2012)). Soient $z = (z_1, \ldots, z_H)$ et $z' = (z'_1, \ldots, z'_{H'})$ deux partitions d'un ensemble $A = \{L_1, \ldots, L_I\}$ et

$w = (w_1, \ldots, w_L)$ et $w' = (w'_1, \ldots, w'_{L'})$ deux partitions d'un ensemble $B = \{C_1, \ldots, C_K\}$. Le principe est de construire un tableau de contingence $\tilde{n} = (\tilde{n}_{p,q})_{(H\times L)\times(H'\times L')}$ tel que $\tilde{n}_{p,q}$ représente le nombre d'éléments de l'ensemble $A \times B$ qui appartiennent à la fois au bloc $p = (h, \ell)$ et au bloc $q = (h', \ell')$. La formule du *Coclustering Adjusted Rand Index* s'écrit

$$CARI((z,w),(z',w')) = \frac{\sum_{p,q}\binom{\tilde{n}^{zwz'w'}_{p,q}}{2} - \sum_{p}\binom{\tilde{n}^{zwz'w'}_{p,\bullet}}{2}\sum_{q}\binom{\tilde{n}^{zwz'w'}_{\bullet,q}}{2}/\binom{I\times K}{2}}{\frac{1}{2}\left[\sum_{p}\binom{\tilde{n}^{zwz'w'}_{p,\bullet}}{2} + \sum_{q}\binom{\tilde{n}^{zwz'w'}_{\bullet,q}}{2}\right] - \left[\sum_{p}\binom{\tilde{n}^{zwz'w'}_{p,\bullet}}{2}\sum_{q}\binom{\tilde{n}^{zwz'w'}_{\bullet,q}}{2}\right]/\binom{I\times K}{2}}.$$

Nous pouvons calculer cet indice de manière très efficace d'un point de vue computationnel car nous avons démontré une relation liant les tableaux de contingence $\tilde{n}^{zwz'w'}$ et $n^{zz'}$, $n^{ww'}$ associés aux *ARI* simples (Hubert and Arabie (1985)) :

$$\tilde{n}^{zwz'w'} = n^{zz'} \otimes n^{ww'},$$

où $\otimes$ représente le produit de Kronecker entre deux matrices.

Application du modèle sur données réelles. Nous appliquons le modèle sur le tableau de contingence issu des données réelles de la base française de pharmacovigilance et collectées entre 2002 et 2010. Dans un premier temps, nous construisons un tableau de contingence réduit en ne considérant que les individus qui ont pris un seul médicament et ont eu seul effet indésirable. Celui-ci contient déjà une information intéressante et nous proposons alors une première liste de signaux potentiels. Ensuite, nous utilisons ces résultats pour effectuer la classification croisée du tableau de contingence général et commentons les résultats obtenus.

Traitement des données individuelles

Les données se présentent sous la forme de deux tableaux binaires. Le premier représente la variable explicative x, réalisation d'une variable aléatoire $X = (X_{ij})_{n\times J}$ et le deuxième tableau, la variable réponse y, réalisation d'une variable aléatoire $Y = (Y_{ik})_{n\times K}$ pour un certain nombre n d'individus définis dans la section 1.1.2.a.

Modèle des blocs latents multiples : principe et estimation des paramètres. Nous étendons le modèle des blocs latents (Govaert and Nadif (2007)) en construisant une partition des lignes $((z, \pi)$ en G groupes et deux partitions des colonnes $((v, \rho)$ et $(w, \tau))$ en H groupes (resp. L groupes), l'une pour les colonnes de x et l'autre pour celles de y (voir figure 1.6). La densité conditionnelle à l'appartenance aux classes des variables x (resp. y) est une loi de Bernoulli de paramètre α (resp. β). Le *modèle des blocs latents multiple*, noté $MLBM$ que nous proposons peut être vu comme un modèle des blocs latents sous contraintes : les classes en colonne ne comportent que des éléments provenant de la même matrice. Il repose sur les hypothèses classiques du modèle des blocs latents mais l'originalité du modèle réside dans l'hypothèse suivante :

$$p(x, y|z, v, w) = p(x|z, v)p(y|z, w).$$

Même si nous supposons une indépendance entre x et y sachant les classes, x et y seront de toute manière liés par la variable latente z. De plus, nous aurons à estimer $G - 1 + H -$

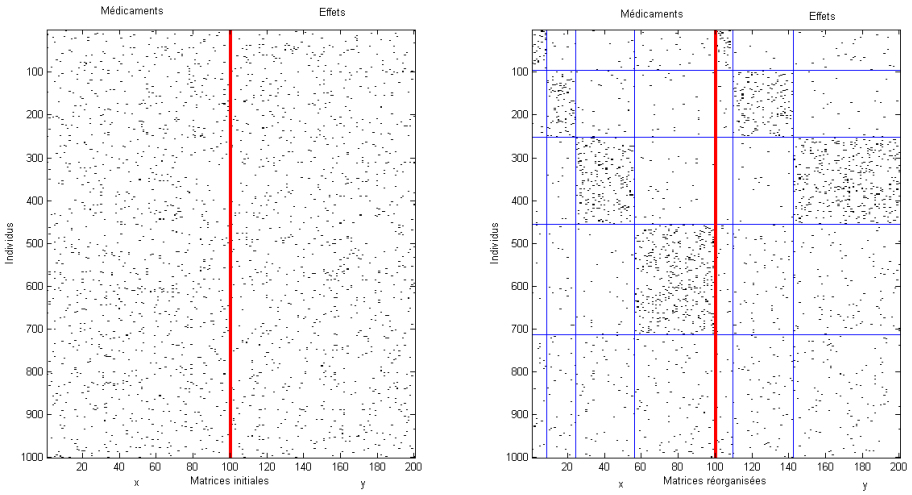


Figure 1.6 – *Matrices simulées x et y de données binaires (à gauche), réorganisées (à droite) avec la partition sur les effets, celle sur les médicaments et celle appariée sur les individus.*

$1+L-1+G\times(H+L)$ paramètres, ce qui est nettement moindre que si nous avions supposé $p(x,y|z,v,w) = p(x|z,v,w)p(y|z,v,w)$.

Pour ce modèle, nous avons identifié les contraintes pour que le modèle soit identifiable en adaptant le théorème de Keribin et al. (2015) :

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 $1 \leq g \leq G$, $\pi_g > 0$ et les coordonnées des vecteurs $\mu = A\rho$ (resp. $\nu = B\tau$) sont distinctes.*

- *C2 : Pour tout $1 \leq h \leq H$ et $1 \leq \ell \leq L$, $\rho_h > 0$ 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 $n \geq \max(2H-1, 2L-1)$ et $min(J,K) \geq 2G-1$.

De même, nous utilisons la procédure d'estimation partiellement bayésienne de Keribin et al. (2015) en mettant les lois a priori suivantes sur les paramètres.

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

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

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

Nous décidons de suivre également leurs recommandations et de choisir $(a,b) = (4,1)$.

Sélection de modèles. Différents critères de sélection de modèles ont été de manière analogue développés et testés sur données simulées (*Integrated Completed Loglikelihood*, ICL

(Biernacki et al. (2000)); *Bayesian Criterion Information*, BIC Schwarz et al. (1978)). De plus, nous étendons la procédure *Bi-KM1* et l'erreur de classification croisée de Lomet (2012) au cadre du modèle développé. La distance de classification entre deux triplets $p = (z, v, w)$ et $q = (z', v', w')$ que nous proposons est la suivante :

$$\text{dist}(p,q) \quad = \quad 1 - \max_{\sigma} \max_{\tau} \max_{\eta} \frac{1}{n(J+K)} \sum_{i,g} z_{ig} z'_{i\sigma(g)} \left(\sum_{j,h} v_{jh} v'_{j\tau(h)} + \sum_{k,\ell} w_{k\ell} w'_{k\eta(\ell')}) \right),$$

où $\sigma \in \mathfrak{S}(\{1,..G\}), \tau \in \mathfrak{S}(\{1,..H\})$ et $\eta \in \mathfrak{S}(\{1,..L\})$.

Un modèle de simulation numérique des données individuelles. Nous proposons un modèle de simulation numérique de la matrice des médicaments x et de la matrice des effets y à partir d'une matrice S de signaux dont nous contrôlons les paramètres comme par exemple le nombre de signaux impliqués. Grâce à ce modèle de simulation, nous mettons en évidence la sensibilité des méthodes existantes face au phénomène de coprescription et nous pouvons ainsi les confronter entre elles et montrer l'intérêt de notre méthodologie vis-à-vis de ce même phénomène. Nous testons également le modèle $MLBM$ sur ces données simulées.

Application aux données réelles. Les données individuelles étant de trop grande taille, nous proposons d'effectuer un sondage pour sélectionner des individus. Nous appliquons alors le modèle sur les données individuelles sélectionnées issues des données réelles de la base française de pharmacovigilance et collectées entre 2002 et 2010. Les résultats précédents obtenus sur le tableau de contingence sont alors utilisés comme information a priori lors du traitement des données individuelles par le modèle $MLBM$. Ceci permettra un premier traitement des données individuelles en prenant en compte l'interaction entre plusieurs effets indésirables.

1.4 Structure du manuscrit

Le manuscrit est organisé de la manière suivante. Le chapitre 2 passe en revue les méthodes de génération automatique de signaux utilisés dans les systèmes officiels de pharmacovigilance et présente une discussion sur celles-ci.

La partie I est consacrée au traitement du tableau de contingence. Dans le chapitre 3, nous reprenons le modèle des blocs latents de Poisson proposé par Govaert and Nadif (2013) et adaptons la procédure d'estimation partiellement bayésienne de Keribin et al. (2015) dans ce cadre. Nous y proposons également une discussion inédite autour des hyperparamètres à partir de données simulées et proposons un indice de classification croisée *Coclustering Adjusted Rand Index* parfaitement adéquat au cadre des données massives (contrairement à l'indice proposé par Lomet (2012)) et adapté du très classique et populaire *Adjusted Rand Index* en classification simple.

Dans le chapitre 4, d'une part, nous présentons les résultats théoriques de deux critères de sélection de modèles, soit ICL et BIC. D'autre part, nous proposons une nouvelle procédure *Bi-KM1* qui permet de parcourir de manière non exhaustive le nombre de classes en ligne et en colonne.

Dans le chapitre 5, le modèle est mis en œuvre pour des données réelles de contingence de pharmacovigilance : estimation des paramètres inconnus et du nombre de classes en ligne et en colonne.

La partie II est consacrée au traitement des données individuelles. Le chapitre 6, nous développons un nouveau modèle ($MLBM$) basé sur le modèle des blocs latents pour traiter

les données individuelles et nous identifions les contraintes assurant l'identifiabilité de ses paramètres. Nous y présentons également la procédure d'estimation utilisée avec une discussion autour des hyperparamètres et proposons une extension de l'erreur de classification proposée par Lomet (2012).

Le chapitre 7 présente les critères ICL et BIC associés à ce modèle. La procédure *Bi-KM1* est adaptée et des tests sur données simulées sont effectués pour évaluer leurs performances.

Enfin, la partie III conclut ce manuscrit. Le chapitre 8 résume l'ensemble du travail réalisé dans cette thèse en proposant une méthode pour analyser les données individuelles de pharmacovigilance. Tout d'abord, nous présentons la méthodologie : dans un premier temps, nous analysons le tableau de contingence "médicaments-effets indésirables" à l'aide du modèle des blocs latents exposé dans la partie I de cette thèse. Cette analyse fournit une liste de groupes de médicaments et de groupes d'effets indésirables pertinentes. À partir de cette liste, nous effectuons une analyse du tableau de données individuelles à l'aide du modèle de blocs latents multiples développés en partie II. Dans un second temps, nous proposons de tester cette méthodologie sur des données simulées. Nous exposons en détail le modèle de simulation numérique utilisé. Puis, nous illustrons cette méthodologie sur un jeu de données réelles non librement accessible : la base française de pharmacovigilance entre 2002 et 2010, administrée par l'AFFSAPS et rendue disponible par l'équipe B2PHI (UMR 1181, INSERM, Villejuif).

Le chapitre 9 présente les conclusions et perspectives de ce travail et évoque les améliorations apportées par cette méthode et les perspectives qu'elle offre.

2

Revue des méthodes en pharmacovigilance

2.1 Les principales méthodes existantes sur le tableau de contingence

Les méthodes de détection automatique travaillant sur les données agrégées en tableau de contingence reposent sur des mesures de disproportionnalité calculées pour l'ensemble des couples présents dans le tableau.

Ainsi, ces dernières sont calculées pour un couple (j, k) en utilisant un tableau de contingence 2×2 résumée du tableau de contingence initial et décrit par la Table 2.1:

	Effet indésirable k	Autres effets	
Médicament j	c_{jk}	$c_{j\bar{k}}$	c_{j+}
Autres médicaments	$c_{\bar{j}k}$	$c_{\bar{j}\bar{k}}$	$c_{\bar{j}+}$
	c_{+k}	$c_{+\bar{k}}$	c_{tot}

Table 2.1 – Tableau de contingence 2×2 pour le couple médicament-effet (j, k).

où c_{jk} représente le nombre de notifications impliquant à la fois le médicament j et l'effet indésirable k, $c_{j+} = \sum_k c_{jk}$, $c_{+k} = \sum_j c_{jk}$ et $c_{tot} = \sum_{j,k} c_{jk}$ représentent les comptes marginaux. Il faut cependant remarquer que ces comptes marginaux ne correspondent pas au nombre de notifications impliquant le médicament j, l'effet indésirable k ou les deux car plusieurs couples sont souvent impliqués dans une seule notification.

2.1.0.a La méthode Reporting Odds Ratio (ROR)

La méthode *Reporting Odds Ratio* (ROR) initiée par van Puijenbroek et al. (2002) vise à estimer pour chacun des couples (j,k) l'odds ratio suivant :

$$\widehat{\mu}_{jk} = \frac{c_{jk}c_{\overline{jk}}}{c_{\overline{jk}}c_{j\overline{k}}}.$$

Le logarithme de $\widehat{\mu}_{jk}$ est supposé suivre une loi normale dont la variance est estimée à partir de la delta méthode. Un signal est alors généré lorsque la borne inférieure de l'intervalle de confiance à 95% de $\log \widehat{\mu}_{jk}$ est strictement supérieure à 0.

Remarque. $\widehat{\mu}_{jk}$ n'est pas calculable dans les rares cas où le médicament j n'est associé qu'à l'effet indésirable k ($c_{j\overline{k}}$=0) ou lorsque l'effet indésirable k n'est associé qu'au médicament j ($c_{\overline{jk}}$=0).

2.1.0.b La méthode Proportionnal Reporting Ratio (PRR)

La méthode *Proportionnal Reporting Ratio* (PRR) initiée par Evans et al. (2001) est basée sur le risque relatif observé pour chacun des couples (j,k) :

$$\widehat{\nu}_{jk} = \frac{\dfrac{c_{jk}}{c_{j+}}}{\dfrac{c_{\overline{jk}}}{c_{\overline{j}+}}}.$$

De manière analogue, $\widehat{\nu}_{jk}$ n'est pas calculable dans le cas où $c_{\overline{jk}} = 0$.

La règle proposée par Evans et al. (2001) pour générer un signal repose sur 3 critères :

- $\widehat{\nu}_{jk} \geqslant 2$,
- $c_{jk} \geqslant 3$,
- la statistique de χ^2 à 1 degré de liberté est supérieure à 4.

Par ailleurs, van Puijenbroek et al. (2002) ont proposé la règle similaire utilisée pour la méthode ROR.

Remarque. Ces deux statistiques $\widehat{\nu_{jk}}$ et $\widehat{\mu_{jk}}$ donnent des résultats très proches, ce qui s'explique par le fait qu'on observe pour la très grande majorité des couples $c_{jk} \ll (c_{\overline{jk}}, c_{j\overline{k}}) \ll c_{\overline{jk}}$ (Almenoff et al. (2007)).

2.1.0.c La méthode Bayesian Confidence Propagation Neural network (BCPNN)

L'approche proposée par Bate et al. (1998) repose sur trois modèles beta-binomiaux :

$$\begin{aligned} c_{jk}|p_{jk} &\sim B(c, p_{jk}) \text{ avec } p_{jk} \sim Be(\alpha_{jk}, \beta_{jk}) \\ c_{j+}|p_{j+} &\sim B(c, p_{j+}) \text{ avec } p_{j+} \sim Be(\alpha_{j+}, \beta_{j+}) \\ c_{+k}|p_{+k} &\sim B(c, p_{+k}) \text{ avec } p_{+k} \sim Be(\alpha_{+k}, \beta_{+k}), \end{aligned}$$

dans lesquels p_{j+}, p_{+k} et p_{jk} désignent respectivement la probabilité dans la base d'être exposé au médicament j, d'observer l'événement indésirable k et de rencontrer les deux. On

peut montrer que les distributions des paramètres a posteriori sont aussi des lois Bêta. Les hyperparamètres quant à eux sont tels que :

$$\alpha_{jk} = 1, \quad \beta_{jk} = \frac{1}{E(p^*_{j+})E(p^*_{+k})} - 1, \quad \alpha_{j+} = 1, \quad \beta_{j+} = 1, \quad \alpha_{+k} = 1 \text{ et } \beta_{+k} = 1,$$

où * est utilisé pour désigner les variables aléatoires conditionnelles aux observations appropriées. Les valeurs choisies pour α_{jk} et β_{jk} sont justifiées par le fait que l'espérance a priori de p_{jk} correspond au produit a posteriori des probabilités marginales.

Remarque. Norén et al. (2006) ont proposé de généraliser ces trois modèles à un modèle Dirichlet-multinomial afin de prendre en compte les dépendances existant entre la probabilité de la cellule considérée et les probabilités des marginales correspondantes.

La statistique d'intérêt est basée sur *Information Component* (IC) défini pour le couple (j, k) qui quantifie la différence entre un nombre attendu $E_{j,k}$ calculé sous l'indépendance des événements j et k et un nombre observé. Plus précisément, on définit IC de la manière suivante :

$$IC^*_{jk} = \log_2 \left(\frac{p^*_{jk}}{p^*_{j+}p^*_{+k}}\right).$$

On peut alors proposer une estimation de cette quantité en posant $E_{jk} = \frac{c_{j+}c_{+k}}{c}$, on a alors

$$IC \approx \log_2 \frac{\frac{c_{jk}}{c}}{\frac{c_{j+}}{c}\frac{c_{+k}}{c}} = \frac{c_{jk}}{E_{jk}} = \frac{c_{jk}}{\frac{(c_{jk}+c_{\bar{j}k})(c_{jk}+c_{j\bar{k}})}{c}}$$

Cependant, comme il y a parfois des événements rares, on lui préfère comme estimation (Caster et al. (2010)) :

$$IC \approx \log_2 \frac{c_{jk} + 0.5}{E_{jk} + 0.5}$$

On prend alors la décision de générer un signal si $Q_{0.025}(IC_{jk}) > 0$.

2.1.0.d La méthode Gamma Poisson Shrinker (GPS)

La méthode proposée par DuMouchel (1999) suppose que les c_{jk} suivent une loi de Poisson :

$$c_{jk} \sim \mathcal{P}(\lambda_{jk}e_{jk}),$$

où e_{jk} est une quantité fixe indiquant le nombre de notifications attendu dans la cellule (j, k) en supposant l'indépendance entre les lignes (médicaments) et les colonnes (effets indésirables) du tableau de contingence : $e_{jk} = \frac{c_{j+}c_{+k}}{c}$. Ensuite, on suppose également le paramètre λ_{jk} suit un mélange de deux lois Gamma :

$$\lambda_{jk} \sim \widehat{w}\,\Gamma(\widehat{\alpha_1}, \widehat{\beta_1}) + (1 - \widehat{w})\Gamma(\widehat{\alpha_2}, \widehat{\beta_2}),$$

où les hyperparamètres sont estimés par maximisation de la vraisemblance marginale des c_{jk} (approche bayésienne empirique).

Remarque. La méthode utilisée par la Food Drug Administration (FDA) propose un calcul plus sophistiqué de e_{jk} en tenant compte des variables comme l'âge ou le sexe des patients. De plus la méthode MGPS permet de s'intéresser à des associations impliquant

plusieurs médicaments ou plusieurs événements indésirables à l'aide de modèles log-linéaires (DuMouchel and Pregibon (2001)).

La règle de décision initialement proposée consistait à ranger les couples (j,k) en fonction de l'espérance a posteriori de $E(\log_2[\lambda])$. Plus récemment, DuMouchel and Pregibon (2001) proposaient une autre possibilité consistant à ranger les cellules selon le quantile à 5% de la distribution des λ^*_{jk} : $Q_{0.05}(\lambda^*_{jk})$. Par la suite, Szarfman et al. (2002) ont proposé le seuil de détection suivant :

$$Q_{0.05}(\lambda^*_{jk}) \geqslant 2.$$

2.2 Vers l'utilisation des données individuelles de pharmacovigilance

Les méthodes basées sur les projections des données en tableau de contingence certes plus simples à traiter d'un point de vue algorithmique, présument d'une certaine homogénéité des individus à l'origine des notifications. Or il est raisonnable de supposer une certaine hétérogénéité dans la population étudiée, d'où l'utilisation des données individuelles de pharmacovigilance. Par ailleurs, nous verrons par la suite, que ces dernières permettent de pallier deux phénomènes rencontrés en utilisant les tableaux de contingence. En effet, il peut y avoir un effet masquant dû par exemple à la forte notification d'un médicament dans le calcul de la mesure de disproportionnalité impliquant un effet, et la coprescription de médicaments qui peut fausser également la mesure de disproportionnalité et attribuer un effet indésirable indu à un médicament. Dans ce qui suit, une méthode prenant en compte les données individuelles est explicitée et les effets de masquage et de coprescription sont ensuite plus détaillés. Le modèle sur lequel est basée cette méthode (Caster et al. (2010)) considère un effet indésirable à la fois. Soit $y = (y_1, \ldots, y_n)$ le vecteur pour lequel chaque coordonnée représente l'absence ou la présence de $\mathcal{Y}$ dans chaque notification et x le vecteur binaire des variables explicatives qui représente la présence ou non de chaque médicament. Le modèle de régression logistique considéré est donc le suivant :

$$\log \frac{P(y|x)}{1-P(y|x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{|p|} x_{|p|}.$$

Les estimations de β_j peuvent être vus comme le log de rapport de côtes ajustés pour l'impact des autres covariables. Ce modèle est étendu à un modèle comprenant une méthode LASSO, c'est-à-dire en imposant une contrainte L1 sur ces coefficients du type :

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Il a par ailleurs été montré que l'estimateur de β sous la contrainte précédente était équivalent au maximum a posteriori (MAP) sous la loi a priori pour chaque β_j la loi de Laplace. Néanmoins, elle semble être très coûteuse d'un point de vue algorithmique.

Il existe également d'autres méthodes basées elles aussi sur la régression logistique mais qui diffèrent soit par une étape de sélection de modèles via le critère Bayesian Information Criterion (BIC) (Marbac et al. (2016)) soit par une étape de présélection manuelle de médicaments (Harpaz et al. (2013)), ou automatique via un rééchantillonnage (Ahmed et al. (2016)). Mais toutes ces méthodes ne considèrent qu'un seul effet indésirable à la fois et le modèle développé dans cette thèse sur les données individuelles (MLBM) permet de remédier à cette problématique (Robert et al. (2015)).

2.3 Tableau récapitulatif

	Type de méthode		Type de données		Limites			Références
	Fréquentiste	Bayésienne	Contingence	Données individuelles	Effet de masquage	Effet de coprescription	prise en compte d'un seul effet indésirable	
Bayesian Confidence Propagation Neural Network		✓	✓		✓	✓		Bate et al. (1998) ; Norén et al. (2006)
(Multi-item) Gamma Poisson Shrinker		✓	✓		✓	✓		DuMouchel (1999) ; DuMouchel and Pregibon (2001)
Proportional Reporting Ratio	✓		✓		✓	✓		Evans et al. (2001)
Reporting Odds Ratio	✓		✓		✓	✓		van Puijenbroek et al. (2002)
Lasso Logistic regression	✓			✓			✓	Caster et al. (2010); Marbac et al. (2016); Harpaz et al. (2013); Ahmed et al. (2016)
LBM avec normalisation		✓	✓			✓		Robert et al. (2016)
MLBM		✓		✓				Robert et al. (2015)

Part I

Traitement du tableau de contingence

3

Modèle des blocs latents (LBM) pour le tableau de contingence

Soit $c = \{c_{jk}; j = 1, \ldots, J; k = 1, \ldots, K\}$, un tableau réalisation d'une variable aléatoire $C = (C_{jk})$. 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. Dans l'exemple représenté en Figure 3.1, la matrice de taille $(J, K) = (200, 400)$ peut être résumée en une matrice $(H, L) = (3, 2)$.

Dans ce but, nous étudions le modèle des blocs latents pour les observations de Poisson développé par Govaert and Nadif (2013). Nous proposons pour les données de comptage, une adaptation de l'algorithme *V-Bayes* qui avait été proposé par Keribin et al. (2015) et étudié

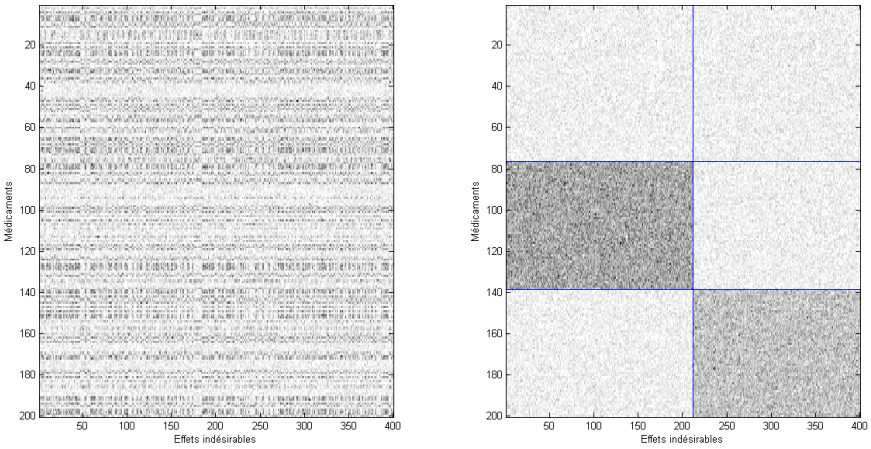


Figure 3.1 – *Matrice simulée c de données de comptage (à gauche), réorganisée (à droite) avec la partition sur les lignes et celle sur les effets.*

dans le cas de données catégorielles. Cette version partiellement bayésienne de l'algorithme VEM Govaert and Nadif (2013) sera l'occasion, dans le cas des données de contingence, de nourrir une discussion inédite autour des hyperparamètres intervenant dans l'inférence bayésienne utilisée.

3.1 Présentation et hypothèses du modèle

Le modèle des blocs latents (LBM, Govaert and Nadif (2007)), repose sur plusieurs hypothèses que nous présentons ici :

- Il existe une structure en blocs des données et ces blocs sont obtenus par le produit cartésien d'une partition des lignes en H composantes représentée par $v = (v_{jh}; j = 1, \ldots, J; h = 1, \ldots, H)$ et d'une partition des colonnes en L composantes représentée par $w = (w_{k\ell}; k = 1, \ldots, K; \ell = 1, \ldots, L)$.
- Les variables V et W sont indépendantes :

$$\forall (v, w) \in \mathcal{V} \times \mathcal{W}, \quad p(v, w) = p(v)p(w),$$

  avec $p(v) = \prod_{j,h} \rho_h^{v_{jh}}$ et $p(w) = \prod_{k,\ell} \tau_\ell^{w_{k\ell}}$, où $(\rho_h = \mathbb{P}(v_{jh} = 1), j = 1, \ldots, J)$ et $(\tau_\ell = \mathbb{P}(w_{k\ell} = 1), \ell = 1, \ldots, L)$ sont les proportions des composantes en ligne et en colonne.
- Les variables aléatoires C_{jk} sont indépendantes conditionnellement à v et w. De plus, ces variables C_{jk} suivent une loi paramétrique notée ϕ qui dépend de la nature des données modélisées. En effet, le LBM peut s'appliquer à plusieurs types de données aux données binaires (Govaert and Nadif (2007)) en utilisant la loi de Bernoulli, aux données réelles en utilisant la loi gaussienne (Lomet (2012)), aux données catégorielles

en utilisant la loi multinomiale (Keribin et al. (2015)), aux données ordinales en utilisant le modèle *BOS* (*Binary Ordinal Search*, Biernacki and Jacques (2015)), aux données fonctionnelles en utilisant la base fournie par l'*ACP* fonctionnel (travail en cours de Ben Slimen et al. (2016) et aux données de comptage (Govaert and Nadif (2013), Aubert et al. (2014)) en utilisant la loi de Poisson. Et nous considérons alors l'approche développée Govaert and Nadif (2013) pour modéliser le tableau de contingence $c = (c_{jk})_{J\times K}$. Ainsi, la distribution paramétrique conditionnelle $\phi(c_{jk}; \mu_j\nu_k\gamma_{h\ell})$ de la variable C_{jk} sachant que v_{jh} et $w_{k\ell}$ valent 1, est supposée être une loi de Poisson $\mathcal{P}(\mu_j\nu_k\gamma_{h\ell})$ où μ_j représente l'effet ligne (de la sorte, deux lignes proportionnelles seront mises dans le même bloc), ν_k représente l'effet colonne et $\gamma_{h\ell}$ représente l'interaction à l'intérieur du bloc $h\ell$. Nous verrons dans la suite, qu'une manière naturelle d'estimer au préalable μ_j et ν_k est de considérer respectivement la marginale $\sum_k c_{jk} = c_{j.}$ et la marginale $\sum_j c_{jk} = c_{.k}$.

La densité conditionnelle pour une observation c_{jk} du bloc $k\ell$ s'écrit alors,

$$\phi(c_{jk}; \mu_j\nu_k\gamma_{h\ell}) = e^{-\mu_j\nu_k\gamma_{h\ell}} \frac{(\mu_j\nu_k\gamma_{h\ell})^{c_{jk}}}{c_{jk}!},$$

Remarque. Nous retrouvons la modélisation $\mathcal{P}(\gamma_{h\ell})$ en prenant $\forall j,k$ μ_j =constante et ν_k =constante. Nous montrerons dans la suite l'intérêt de la normalisation sur des données simulées.

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

$$p(c;\theta) = \sum_{(v,w)\in\mathcal{V}\times\mathcal{W}} p(v;\theta)p(w;\theta)p(c|v,w;\theta) \tag{3.1}$$

$$= \sum_{(v,w)\in\mathcal{V}\times\mathcal{W}} \prod_{j,h} \rho_h^{v_{jh}} \prod_{k,\ell} \tau_\ell^{w_{k\ell}} \prod_{h,j,k,\ell} \phi(c_{jh}; \mu_j\nu_k\gamma_{h\ell})^{v_{jh}w_{k\ell}}, \tag{3.2}$$

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

3.2 Estimation des paramètres

Remarquons qu'il n'est pas possible de factoriser les termes dans (3.2) pour contourner le calcul de la somme sur chaque couple. Pour effectuer l'estimation des paramètres du modèle des blocs latents, on considère alors la vraisemblance complétée des données c. Les partitions v, w ne sont pas observées et sont donc des variables latentes. L'algorithme *Espérance Maximisation* (*EM*), introduit par Dempster et al. (1977), est un algorithme classique et incontournable pour l'estimation des paramètres d'un modèle comportant des variables latentes. Après une étape d'initialisation des paramètres du modèle θ_0, cet algorithme alterne deux étapes successives à chaque itération (d) :

- **Espérance** : La première étape consiste à effectuer la maximisation de l'espérance de la log vraisemblance complétée conditionnellement aux observations c et aux paramètres courants estimés notés $\theta_{H,L}^{(d)}$

$$\mathcal{Q}(\theta_{H,L}|\theta_{H,L}^{(d)}) = \mathbb{E}\left[\log \ell(\theta_{H,L}; v, w)|c, \theta_{H,L}^{(d)}\right].$$

Pour les modèles de mélange, cette étape revient à calculer les probabilités conditionnelles, notées $t_{jh}^{(d)}$ et $r_{k\ell}^{(d)}$ que l'observation c_{jk} appartient à la composante h en ligne et ℓ en colonne sachant les données c et les paramètres courant du modèle $\theta_{H,L}^{(d)}$

- **Maximisation** : La deuxième étape de l'algorithme consiste à trouver $\theta_{H,L}^{(d)}$, les paramètres maximisant la quantité $\mathcal{Q}(\theta_{H,L}|\theta_{H,L}^{(d)})$ en $\theta_{H,L}$. La propriété fondamentale de cet algorithme est la suivante : la maximisation de la quantité $\mathcal{Q}(\theta_{H,L}|\theta_{H,L}^{(d)})$ garantit d'augmenter la log vraisemblance $\ell(c|\theta_{H,L})$.

Ainsi, l'algorithme EM alterne les étapes d'Espérance et de Maximisation et met à jour les paramètres du modèle à chaque itération (d) jusqu'à convergence, parfois lente, vers un maximum local. Celui-ci n'est d'ailleurs pas forcément le maximum global de la fonction de vraisemblance.

Néanmoins, si nous reprenons le calcul de l'espérance conditionnellement aux observations et sachant $\theta^{(d)}$ de la log-vraisemblance complétée utilisé dans l'algorithme EM, nous avons :

$$\mathcal{Q}(\theta;\theta^{(d)}) = \sum_{v,w} \log p(c,v,w;\theta) p(v,w|c;\theta^{(d)})$$

et

$$p(c,v,w;\theta) = p(c|v,w;\theta)p(v,w;\theta).$$

D'où

$$\mathcal{Q}(\theta;\theta^{(d)}) \quad = \quad \sum_{j,h} r_{jh}^{(d)} \log \rho_h + \sum_{k,\ell} t_{k\ell}^{(d)} \log \tau_\ell + \sum_{j,h,k,\ell} e_{jhk\ell}^{(d)} \log \phi(c_{jk};\mu_j\nu_k\gamma_{k\ell})$$

où $\tilde{r}_{jh}^{(d)} = \mathbb{P}(V_{jh}=1|c;\theta^{(d)})$, $\tilde{t}_{k\ell}^{(d)} = \mathbb{P}(W_{k\ell}=1|c;\theta^{(d)})$, $e_{jkh\ell}^{(d)} = \mathbb{P}(V_{jh}=1, W_{k\ell}=1|c;\theta^{(d)})$.

Le calcul de $e_{jkh\ell}$ n'est pas réalisable en un temps fini raisonnable car nous n'avons pas l'indépendance des variables latentes conditionnellement aux observations. Pour contourner cette difficulté, Govaert and Nadif (2008) proposent l'algorithme VEM qui propose de faire cette approximation d'indépendance conditionnelle qui permet d'approcher l'étape E de l'algorithme EM.

3.2.1 L'algorithme *Variational Expectation Maximisation* (VEM)

Le principe est d'écrire la log-vraisemblance en introduisant une distribution libre quelconque des variables latentes $q_{vw}(v,w)$. Nous pouvons alors décomposer la log-vraisemblance comme la somme de deux fonctions :

$$L(\theta) = \underbrace{\mathbb{E}_{(V,W)\sim q_{vw}}\left[\log\left(\frac{p(c,V,W;\theta)}{q_{vw}(V,W)}\right)\right]}_{:=\mathcal{F}(q_{vw};\theta)} + D_{KL}\big(q_{vw}||p(.,.|c;\theta)\big).$$

où $\mathcal{F}$ est une fonction de la distribution libre q_{vw} ou *énergie libre* et D_{KL} la divergence de Kullback-Leibler. Comme la divergence de Kullback-Leiber est positive, l'*énergie libre* est un minorant de la log-vraisemblance et égale à cette dernière si et seulement si la distribution libre q_{vw} est égale à $p(v,w|c;\theta^{(d)})$.

Ainsi, calculer la loi $p(v, w|c; \theta^{(d)})$ dans l'étape E revient à maximiser l' énergie libre $\mathcal{F}(q_{vw})$ en q_{vw}, si la fonction q_{vw} est recherchée parmi l'ensemble des lois possibles.

Quand la structure de covariance en v et w est trop compliquée, on peut rechercher une approximation de $p(v, w|c; \theta^{(d)})$ parmi les lois q_{vw}. L'idée est alors de faire une approximation variationnelle ou dite en champ moyen, c'est-à-dire de maximiser l'énergie libre à $\theta^{(d)}$ fixé en utilisant en supposant que les distributions libres q_{vw} se factorisent :

$$q_{vw}(v, w) = q_v(v)q_w(w).$$

L'avantage de cette simplification est qu'elle permet de calculer facilement les mises à jour de la loi. Ainsi, l'algorithme *Variational Expectation Maximisation* (VEM) s'écrit :

$$\begin{aligned}\mathcal{F}(q_{vw}; \theta) &= \sum_{j,h} r_{jh}^{(d)} \log \rho_h + \sum_{k,\ell} t_{k\ell}^{(d)} \log \tau_\ell + \sum_{j,k,h,\ell} r_{jh}^{(d)} t_{k\ell}^{(d)} \log \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell}) \\ &\quad - \sum_{j,h} r_{jh}^{(d)} \log r_{jh}^{(d)} - \sum_{k,\ell} t_{k\ell}^{(d)} \log t_{k\ell}^{(d)}.\end{aligned}$$

Remarquons que l'énergie libre est un minorant de la log-vraisemblance dont la qualité dépend de la divergence de Kullback-Leiber qui ne peut être estimée. Les deux problèmes qui se posent sont de savoir si

- le paramètre qui maximise l'énergie libre est proche de celui qui maximise la log-vraisemblance,
- la valeur maximale de l'énergie libre est proche de celle de la log-vraisemblance.

Ces deux questions trouvent leur réponses asymptotiquement dans le cas de la modélisation sans normalisation, grâce aux résultats théoriques présentés dans l'introduction (section 2.1.3 b.) sur la normalité asymptotique de l'EMV et du maximum de l'énergie libre.

L'algorithme VEM est donc le suivant :

Algorithme VEM pour données de comptage.

1. Initialisation de $\theta^{(0)}$ et de $t_{k\ell}^{(0)}$.
2. Pour $d = 0 \dots n_{\text{iter}}$:
   - Étape VE : maximisation alternée de l'énergie libre à $\theta^{(d)}$ fixé en prenant $t_{k\ell}^{(t=0)} = t_{k\ell}^{(d)}$:
     - calcul de $r_{jh}^{(e+1)}$ à $t_{k\ell}^{(e)}$ et à $\theta^{(d)}$ fixés:

$$r_{jh}^{(e+1)} = \frac{\rho_h^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell}^{(d)} t_{k\ell}^{(e)}} \prod_\ell \left(\gamma_{h\ell}^{(d)}\right)^{\sum_k t_{k\ell}^{(e)} c_{jk}}}{\sum_{h'} \rho_{h'}^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell}^{(d)} t_{k\ell}^{(e)}} \prod_\ell \left(\gamma_{h'\ell}^{(d)}\right)^{\sum_k t_{k\ell}^{(e)} c_{jk}}}$$

     - calcul de $t_{k\ell}^{(e+1)}$ à $r_{jh}^{(e+1)}$ et à $\theta^{(d)}$ fixés.

   $\longrightarrow$ Obtention des probabilités $r_{jh}^{(d+1)}$ et $t_{k\ell}^{(d+1)}$.

   - Étape M : calcul du paramètre $\theta^{(d+1)}$:

$$\rho_h^{(d+1)} = \frac{\sum_{j=1}^{J} r_{jh}^{(d+1)}}{J}, \quad \rho_h^{(d+1)} = \frac{\sum_{k=1}^{K} t_{k\ell}^{(d+1)}}{K} \quad \gamma_{h\ell}^{(d+1)} = \frac{\sum_{j=1}^{J} \sum_{h=1}^{H} r_{jh}^{(d+1)} t_{k\ell}^{(d+1)} c_{jk}}{\sum_{j=1}^{J} \mu_j r_{jh}^{(d+1)} \sum_{k=1}^{K} \nu_k t_{k\ell}^{(d+1)}}.$$

3. Obtention d'un estimateur $\widehat{\theta}^{VEM} = \theta^{(n_{\text{iter}})}$.

Le critère d'arrêt considéré est le minimum entre *niter* et l'itération pour laquelle l'énergie libre n'évolue plus à un seuil près. Enfin, l'estimation des labels s'effectue en utilisant la règle du Maximum A Posteriori (MAP, voir l'introduction).

L'algorithme VEM par rapport à l'algorithme EM est donc implémentable en pratique. Cependant, l'inconvénient est qu'un point stationnaire de cet algorithme ne peut être un point stationnaire de la vraisemblance que si le modèle satisfait aux conditions de simplification de l'approximation variationnelle (Keribin et al. (2010)); ce qui est parfois réalisé dans certaines conditions asymptotiques, mais qui ne l'est en général pas à distance finie. De plus, cet algorithme qui est déterministe, est encore plus sensible à l'initialisation par rapport à l'algorithme EM.

Une autre alternative à l'algorithme VEM, où n'est effectuée aucune approximation est l'algorithme SEM-Gibbs proposé par Keribin et al. (2010) pour le modèle des blocs latents. L'étape d'estimation est remplacée par la génération d'un échantillon des données manquantes $(v^{(d)}, w^{(d)})$ sous la loi des données manquantes conditionnellement aux observations et au paramètre courant $\theta^{(d)}$: on obtient ainsi un pseudo-échantillon complet (étape S). L'étape de maximisation recherche le paramètre maximisant la vraisemblance complétée, dans laquelle les variables manquantes sont remplacées par leur tirage. Pour contourner l'impossibilité du calcul de la loi $p(v, w|c, \theta^{(d)})$, un échantillonneur de Gibbs (voir section 4.2.3) est utilisé en effectuant une simulation itérative de v suivant $p(w|c, v; \theta^{(d)}$, puis de V suivant $p(v|c; \theta^{(d)}$. L'algorithme simule alors une chaîne de Markov irréductible avec une unique distribution stationnaire qui est concentrée autour de l'estimateur du maximum de vraisemblance.

Toutefois, même si l'algorithme SEM-Gibbs est moins sensible aux initialisations que l'algorithme VEM, Brault (2014) a mis en évidence un nouveau type d'états absorbants pour la chaîne de Markov engendrée par l'algorithme propre à la configuration en blocs du

modèle qui empêchent l'ergodicité de la chaîne obtenue.

Pour pallier tous ces inconvénients et tirer profit de chacun des deux algorithmes présentés, un couplage d'algorithmes SEM-Gibbs+VEM peut être envisagé puisque l'algorithme SEM-Gibbs fournirait en amont une bonne initialisation pour l'algorithme VEM. Néanmoins, Keribin et al. (2015) ont montré que cette combinaison estime parfois un nombre de classes inférieur à celui demandé (phénomène dit de dégénérescence des classes).

Pour contourner cette difficulté, Keribin et al. (2015) ont proposé l'algorithme *V-Bayes* (voir Figure 3.2), version partiellement bayésienne de l'algorithme VEM. Ils ont également montré que le couplage avec l'échantillonneur de Gibbs en amont, est la meilleure combinaison dans le cadre du modèle des blocs latents sur données catégorielles. En effet, l'échantillonneur de Gibbs, tout comme l'algorithme SEM-Gibbs, permet de fournir une zone pertinente autour du bon mode a posteriori et présente l'avantage supplémentaire suivant : les états qui sont absorbants pour l'algorithme SEM-Gibbs ne le sont pas pour lui (Brault (2014)). Une étude plus détaillée et qui a nourri toute cette réflexion sur la comparaison de tous ces algorithmes est proposée dans Keribin et al. (2015).

Dans ce qui suit, nous présentons ces deux algorithmes bayésiens et leurs mises en œuvre dans le cas des données de comptage.

3.2.2 Inférence bayésienne

Nous adaptons l'algorithme *V-Bayes* (Keribin et al. (2015)) qui est basé sur une inférence bayésienne (voir Figure 3.2) 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 :

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

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

$$\gamma_{h\ell}|\delta, \zeta \sim \Gamma(\delta, \zeta).$$

Par ailleurs, nous ne considérons pas ici μ_j et ν_k comme des variables aléatoires et ils sont alors estimés comme précédemment dit, par $c_{j.}$ et $c_{.k}$.

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

$$\sum_j \mu_j = \sum_k \nu_k = \frac{1}{\sum_h \rho_h \gamma_{h\ell}} = \frac{1}{\sum_\ell \tau_\ell \gamma_{h\ell}} = \sum_{j,k} c_{jk}, \forall k, \ell$$

assure les égalités suivantes,

$$\mathbb{E}(\sum_k c_{jk}) = \mu_j \quad \text{et} \quad \mathbb{E}(\sum_j c_{jk}) = \nu_k,$$

d'où l'estimation naturelle proposée pour μ_j et ν_k. En effet, nous avons

$$\begin{aligned}
\mathbb{E}(\sum_k c_{jk}) &= \sum_k \sum_{h,\ell} \mathbb{E}(c_{jk}|v_{jh}=1, w_{k\ell}=1)\mathbb{P}(v_{jh}=1, w_{k\ell}=1) \\
&= \sum_k \sum_{h,\ell} \mu_j \, \nu_k \, \gamma_{h\ell} \, \rho_h \, \tau_\ell \\
&= \mu_j \sum_k \nu_k \sum_h \rho_h \sum_\ell \gamma_{h\ell} \, \tau_\ell \\
&= \mu_j \cancel{\sum_k \nu_k} \sum_h \rho_h \frac{1}{\cancel{\sum_k \nu_k}} \\
&= \mu_j .
\end{aligned}$$

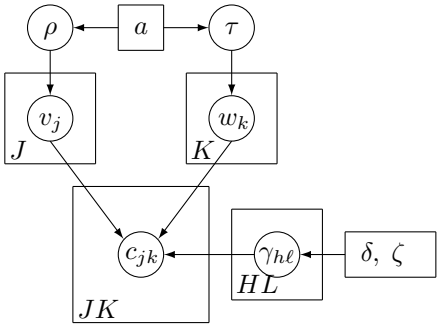


Figure 3.2 – Graphe bayésien du modèle.

Une discussion sur le choix des hyperparamètres est présente après la description suivante des deux algorithmes bayésiens considérés.

3.2.2.a Échantillonneur de Gibbs

L'échantillonneur de Gibbs est une forme particulière de la méthode de Monte-Carlo par chaîne de Markov qui, du fait de sa simplicité, est largement utilisé dans de nombreux domaines d'analyse statistique bayésienne. Cette méthode a été initialement utilisée par Geman and Geman (1984) pour générer des observations à partir d'une distribution de Gibbs (distribution de Boltzmann).

Ainsi, le principe de l'échantillonnage consiste à considérer un paramètre inconnu sous la forme vecteur ayant plusieurs composantes et à échantillonner à partir de la distribution conditionnelle d'une composante quand toutes les autres sont fixées. Dans la méthode de Gibbs, après avoir choisi un point départ, les composantes du vecteur (v, w, θ) sont générées les unes après les autres conditionnellement à toutes les autres composantes. Si $p(v, w, \theta|c)$ est la densité recherchée, conditionnellement aux données observées (c), les densités conditionnelles $p(v|c, w; \theta)$, $p(w|c, v; \theta)$ sont alors utilisées, et ainsi de suite. À chaque d-ième étape, la distribution conditionnelle utilise les valeurs générées les plus récentes parmi toutes les autres composantes. Les itérations successives de cet algorithme génèrent successivement les états d'une chaîne de Markov admettant une mesure invariante qui est la loi a posteriori. Pour un nombre d'itérations suffisamment grand, le vecteur obtenu peut donc être considéré comme étant une réalisation de la loi a posteriori $p(v, w, \theta|c)$.

Le schéma numérique de l'algorithme est décrit ci-dessous dans le cadre du modèle des blocs latents pour données de contingence.

Échantillonneur de Gibbs.

Obtention d'une chaîne de Markov de loi stationnaire $p(v, w, \theta|c)$.

1. Initialisation de $\theta^{(0)}$ et de $w^{(0)}$.
2. Pour $d = 0 \dots n_{\text{iter}}$:
   - Simulation de $v^{(d+1)}$ suivant la loi $p(v|c, w^{(d)}; \theta^{(d)})$.
   - Simulation de $w^{(d+1)}$ suivant la loi $p(w|c, v^{(d+1)}; \theta^{(d)})$.
   - Simulation de $\theta^{(d+1)}$ suivant la loi $p(\theta|c, v^{(d+1)}, w^{(d+1)})$.
3. Obtention d'un estimateur $\widehat{\theta^G} = \frac{1}{n_{iter}} \sum_{d=1}^{n_{iter}} \theta^{(d)}$.

Plus précisément, les lois ci-dessus peuvent être calculées de manière explicite :

- Simulation de $v^{(d+1)}$ sous la loi $p(v|c, w^{(d)}; \theta^{(d)})$:

$$\mathbb{P}(v_j^{(d+1)} = h|w^{(d)}, c, \theta^{(d)}) = \frac{\rho_h^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell}^{(d)} w_{k\ell}^{(d)}} \prod_\ell \left(\gamma_{h\ell}^{(d)}\right)^{\sum_k w_{k\ell}^{(d)} c_{jk}}}{\sum_{h'} \rho_{h'}^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h'\ell}^{(d)} w_{k\ell}^{(d)}} \prod_\ell \left(\gamma_{h'\ell}^{(c)}\right)^{\sum_k w_{k\ell}^{(d)} c_{jk}}}$$

- Simulation de $w^{(d+1)}$ sous la loi $p(w|c, v^{(d+1)}; \theta^{(d)})$:

$$\mathbb{P}(w_k^{(d+1)} = \ell|v^{(d+1)}, c, \theta^{(d)}) = \frac{\tau_\ell^{(d)} e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell}^{(d)} v_{jh}^{(d+1)}} \prod_h \left(\gamma_{h\ell}^{(d)}\right)^{\sum_j v_{jh}^{(d+1)} c_{jk}}}{\sum_{\ell'} \tau_{\ell'}^{(d)} e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell'}^{(d)} v_{jh}^{(d+1)}} \prod_h \left(\gamma_{h\ell'}^{(d)}\right)^{\sum_j v_{jh}^{(d+1)} c_{jk}}}$$

- Tirage de $\rho^{(d+1)}$ suivant la loi $\rho|v^{(d+1)} \sim \mathcal{D}\left(v_{\cdot 1}^{(d+1)} + a, \dots, v_{\cdot H}^{(d+1)} + a\right)$ avec

$$v_{\cdot h}^{(d+1)} = \sum_j v_{jh}^{(d+1)}.$$

- Tirage de $\tau^{(d+1)}$ suivant la loi $\tau|w^{(d+1)} \sim \mathcal{D}\left(w_{\cdot 1}^{(d+1)} + a, \dots, w_{\cdot L}^{(d+1)} + a\right)$ avec

$$w_{\cdot \ell}^{(d+1)} = \sum_k w_{k\ell}^{(d+1)}.$$

- Tirage de $\gamma_{h\ell}^{(d+1)}$ suivant la loi

$$\gamma_{h\ell}|v^{(d+1)}, w^{(d+1)}, c \quad \sim \quad \Gamma\left(\delta + \sum_{j,k} v_{jh} w_{k\ell} c_{jk}, \zeta + \sum_{j,k} v_{j\cdot h}^{(d+1)} w_{k\ell}^{(d+1)} \mu_j \nu_k\right).$$

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 $v_j^{(d)}, w_k^{(d)}$, simulés au cours des itérations.

3.2.2.b Algorithme V-Bayes

Pour estimer le mode de la loi a posteriori de θ, McLachlan and Krishnan (2008) proposent la même démarche que celle de l'algorithme EM :

$$\begin{aligned} \log p(c;\theta) &= \mathcal{Q}(\theta|\theta^{(d)}) - H(\theta|\theta^{(d)}) + \log p(\theta) \\ &= \mathcal{F}_B(\theta|\theta^{(d)}) - H(\theta|\theta^{(d)}), \end{aligned}$$

où $p(\theta)$ désigne la loi a priori de θ définie dans la section 4.2.2 a.

Ainsi, l'algorithme *V-Bayes* cherche à maximiser une version "bayésienne" de l'énergie libre $\mathcal{F}_B$ définie par :

$$\mathcal{F}_B(\theta) = \mathcal{F}(\theta) + \log p(\theta).$$

Le schéma de simulation de l'algorithme est le suivant et a été adapté pour les données de comptage :

Algorithme *V-Bayes* pour données de comptage.

1. Initialisation de $\theta^{(0)}$, de $r_{jh}^{(0)}$ et de $t_{k\ell}^{(0)}$.

2. Pour $d = 0 \dots n_{\text{iter}}$:

   - Étape VE : maximisation alternée de l'énergie libre à $\theta^{(d)}$ fixé en prenant $r_{jh}^{(e=0)} = r_{jh}^{(d)}$ et $t_{k\ell}^{(e=0)} = t_{k\ell}^{(d)}$:

     - calcul de $r_{jh}^{(e+1)}$ à $t_{k\ell}^{(e)}$ et à $\theta^{(d)}$ fixés:

$$r_{jh}^{(e+1)} = \frac{\rho_h^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell}^{(d)} t_{k\ell}^{(e)}} \prod_\ell \left(\gamma_{h\ell}^{(d)}\right)^{\sum_k t_{k\ell}^{(e)} c_{jk}}}{\sum_{h'} \rho_{h'}^{(d)} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell}^{(d)} t_{k\ell}^{(e)}} \prod_\ell \left(\gamma_{h'\ell}^{(d)}\right)^{\sum_k t_{k\ell}^{(e)} c_{jk}}}$$

     - calcul de $t_{k\ell}^{(e+1)}$ à $r_{jh}^{(e)}$ et à $\theta^{(d)}$ fixés:

$$t_{k\ell}^{(e+1)} = \frac{\tau_\ell^{(d)} e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell}^{(d)} r_{jh}^{(e)}} \prod_h \left(\gamma_{h\ell}^{(d)}\right)^{\sum_j r_{jh}^{(e)} c_{jk}}}{\sum_{\ell'} \tau_{\ell'}^{(d)} e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell}^{(d)} r_{jh}^{(e)}} \prod_h \left(\gamma_{h\ell'}^{(d)}\right)^{\sum_j r_{jh}^{(e)} c_{jk}}}$$

   $\longrightarrow$ Obtention des probabilités $r_{jh}^{(d+1)}$ et $t_{k\ell}^{(d+1)}$.

   - Étape M : calcul du paramètre $\theta^{(d+1)}$:

$$\rho_h^{(d+1)} = \frac{a - 1 + \sum_j r_{jh}^{(d+1)}}{J + H(a-1)}, \quad \tau_\ell^{(d+1)} = \frac{a - 1 + \sum_k t_{k\ell}^{(d+1)}}{K + L(a-1)}$$

$$\gamma_{h\ell} = \frac{\delta - 1 + \sum_{j,k} r_{jh}^{(d+1)} t_{k\ell}^{(d+1)} c_{jk}}{\zeta + \sum_{j,k} \mu_j \nu_k r_{jh}^{(d+1)} t_{k\ell}^{(d+1)}}$$

3. Obtention d'un estimateur $\widehat{\theta}^{VB} = \theta^{(n_{\text{iter}})}$.

Remarques importantes.

1. Dans l'étape VE, Govaert and Nadif (2008) ont montré qu'une seule étape alternée (e) suffit pour l'algorithme VEM et nous utilisons ce constat pour l'algorithme *V-Bayes* également.

2. Nous pouvons remarquer que les formules de mises à jour dans l'étape M correspondent aux formules de mises à jour de l'algorithme VEM pour les hyperparamètres $a = 1$, $\delta = 1$, et $\zeta = 0$.

3.2.3 Discussion autour des hyperparamètres

Le choix des hyperparamètres a, δ, ζ reste un problème important en inférence bayésienne. Dans le cas de mélanges gaussiens simples, Frühwirth-Schnatter (2011) propose de prendre des valeurs de a valant 4 ou 16 et ce choix testé empiriquement permet aux algorithmes $MCMC$ de produire moins de classes vides. Keribin et al. (2015) confirmeront le choix de prendre $a = 4$ dans le cas de données catégorielles pour le modèle des blocs latents. Nous

choisissons de suivre cette recommandation pour le choix de la valeur de a, car une valeur de a trop petite a tendance à vider les classes, ce qui est confirmé par ce qui suit.

Très récemment, de nouvelles approches telles que les modèles de mélanges parcimonieux proposée par Malsiner-Walli et al. (2016) choisissent une loi a priori parcimonieuse, soit un paramètre a très petit. Le but de ces méthodes est de surapprendre le nombre de composantes d'un modèle de mélange grâce à une loi a priori parcimonieuse qui videra les classes superflues. L'estimateur du nombre de composantes est alors le nombre le plus fréquent de composantes non vides observées pendant l'échantillonnage des algorithmes $MCMC$. Cette idée est également présente dans la procédure de Wyse et al. (2016) à laquelle nous allons nous comparer dans le chapitre 7.

En ce qui concerne les hyperparamètres δ, ζ, il n'existe pas de loi a priori non informative (voir Figure 3.3) , contrairement à n'importe quelle loi définie sur un espace compact, ce qui peut rendre l'étude plus délicate.

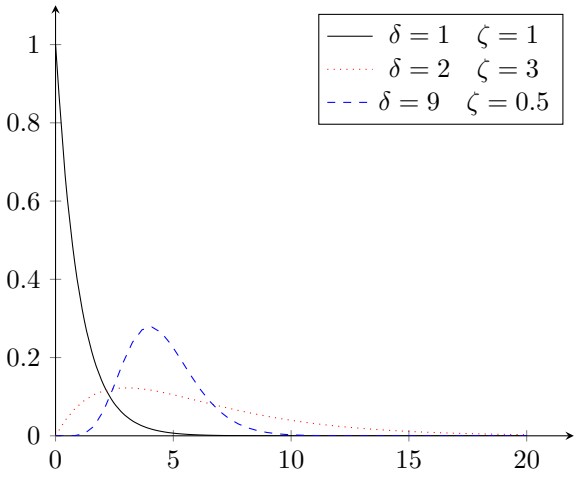


Figure 3.3 – Différentes densités de la distribution $\Gamma(\delta, \zeta)$ selon les valeurs de δ et ζ

3.2.3.a Choix empirique de l'hyperparamètre a

Dans cette partie, nous expérimentons le choix des valeurs de l'hyperparamètre a. Pour illustrer et conforter l'étude déjà effectuée, nous proposons un premier plan d'expérience mettant en jeu des matrices très simples à classifier et nous le complexifierons pour le choix des hyperparamètres δ et ζ, car à notre connaissance, aucune étude pour ces hyperparamètres n'a été faite.

Ainsi, le plan d'expérience consiste à simuler des matrices de comptage avec $H = 4$ et $L = 5$ classes en ligne et colonne, des paramètres pour les blocs de la forme :

$$\gamma = \begin{pmatrix} 15 & 10 & 5 & 5 & 5 \\ 10 & 15 & 10 & 5 & 5 \\ 5 & 10 & 15 & 10 & 5 \\ 5 & 5 & 10 & 15 & 10 \end{pmatrix}.$$

De plus, nous faisons varier :

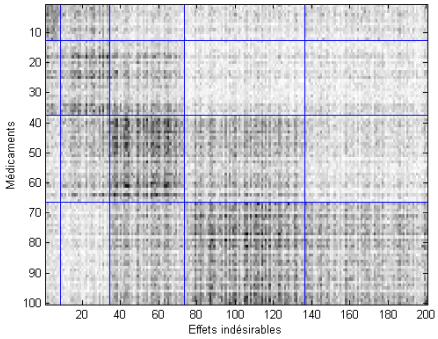


Figure 3.4 – *Exemple de matrice simulée c de données de comptage de tailles* $(100, 200)$ *et réorganisée pour le scénario considéré.*

- les proportions :
  - proportion équilibrée

$$\rho = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \text{ et } \tau = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$$

  - proportion arithmétique

$$\rho = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{pmatrix} \text{ et } \tau = \begin{pmatrix} 0.1 \\ 0.15 \\ 0.2 \\ 0.25 \\ 0.3 \end{pmatrix}$$

- le nombre (J, K) de lignes et de colonnes variant entre 100 et 200.

Ainsi, nous utilisons l'algorithme *V-Bayes* pour des paramètres δ et ζ valant $(10^{-10}, 0.001, 0.01, 0.1, 1, 10)$. Pour chaque configuration, nous simulons 100 matrices. Comme l'algorithme *V-Bayes* est sensible aux initialisations, pour chaque matrice nous générons aléatoirement 10 initialisations et nous lançons avec chacune l'algorithme *V-Bayes*.

Dans les tableaux ci-dessous (voir Figure 3.5 et Figure 3.6), nous recensons le pourcentage de solutions renvoyées par l'algorithme *V-Bayes* ayant au moins une classe vide en ligne ou en colonne.

proportions équilibrées

$a = 0.1$

(100,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	93.9	93.9	93.9	93.9	93.9	93.9
10^{-3}	93.9	93.9	93.9	93.9	93.9	93.9
0.01	93.9	93.9	93.9	93.9	93.9	93.9
0.1	93.9	93.9	93.9	93.9	93.9	93.9
1	51.6	86.1	88.2	89.3	90.3	90.8
10	67.1	66.8	66.8	66.8	66.8	66.8

(200,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	96.4	96.4	96.4	96.4	96.4	96.4
10^{-3}	96.4	96.4	96.4	96.4	96.4	96.4
0.01	96.5	96.5	96.5	96.5	96.5	96.5
0.1	96.4	96.4	96.4	96.4	96.4	96.4
1	55.6	88.4	89.2	90.9	91.8	93.4
10	74.5	74.3	74.3	74.3	74.3	74.2

(200,100)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	97.8	97.8	97.8	97.8	97.8	97.8
10^{-3}	97.8	97.8	97.8	97.8	97.8	97.8
0.01	97.8	97.8	97.8	97.8	97.8	97.8
0.1	97.8	97.8	97.8	97.8	97.8	97.8
1	57.4	91.4	93.3	94.3	94.7	95.6
10	67.4	66.4	66.3	66.2	66.2	66

$a = 4$

(100,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	64.9	64.9	64.9	64.9	64.9	64.9
10^{-3}	64.9	64.9	64.9	64.9	64.9	64.9
0.01	64.9	64.9	64.9	64.9	64.9	64.9
0.1	64.2	64.2	64.2	64.2	64.2	64.2
1	0.2	0.7	0.9	1.7	4.8	8.3
10	36.8	36.7	36.4	36.3	36.5	36.2

(200,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	81	81.3	81.1	81.2	81.3	81.3
10^{-3}	81.2	81.3	81.3	81.3	81.3	81.3
0.01	81.3	81.3	81.3	81.3	81.3	81.3
0.1	80.7	80.4	80.7	80.7	80.7	80.7
1	0.8	4.5	7.6	11.2	16.9	24
10	55.7	55.7	55.7	55.6	55.4	54.8

(200,100)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	78.4	78.3	78.4	78.4	78.4	78.4
10^{-3}	78.3	78.3	78.3	78.3	78.3	78.2
0.01	78.3	78.3	78.3	78.3	78.2	78.3
0.1	77.7	77.7	77.7	77.7	77.6	77.7
1	0.1	2.5	4.2	7.8	13.5	20.9
10	50.2	49.8	49.7	49.6	48.8	47.9

Figure 3.5 – Pourcentages de classes vides obtenues par l'algorithme *V-Bayes* initialisés 10 fois avec $(H, L) = (4, 5)$ classes pour 100 matrices avec proportions équilibrées pour différentes valeurs de a (en ligne) et pour tailles d'échantillons (en colonne).

proportions non équilibrées

$a = 0.1$

(100,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	98.7	98.7	98.7	98.7	98.7	98.7
10^{-3}	98.7	98.7	98.7	98.7	98.7	98.7
0.01	98.7	98.7	98.7	98.7	98.7	98.7
0.1	98.6	98.6	98.6	98.6	98.6	98.6
1	82.9	98.2	98.1	98.4	98.4	98.5
10	90.1	89.5	89.4	89.4	89.4	89.4

(200,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	99.4	99.4	99.4	99.4	99.4	99.4
10^{-3}	99.4	99.4	99.4	99.4	99.4	99.4
0.01	99.4	99.4	99.4	99.4	99.4	99.4
0.1	99.4	99.4	99.4	99.4	99.4	99.4
1	83	98.5	98.7	99.1	99.1	99.1
10	90.1	89.6	89.6	89.6	89.6	89.6

(200,100)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	99.4	99.4	99.4	99.4	99.4	99.4
10^{-3}	99.4	99.4	99.4	99.4	99.4	99.4
0.01	99.4	99.4	99.4	99.4	99.4	99.4
0.1	99.4	99.4	99.4	99.4	99.4	99.4
1	83.8	99.2	99.2	99.2	99.3	99.4
10	90.3	90	90	89.9	89.9	89.7

$a = 4$

(100,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	64.9	64.9	64.9	64.9	64.9	64.9
10^{-3}	64.9	64.9	64.9	64.9	64.9	64.9
0.01	64.9	64.9	64.9	64.9	64.9	64.9
0.1	64.2	64.2	64.2	64.2	64.2	64.2
1	0.2	0.7	0.9	1.7	4.8	8.3
10	36.8	36.7	36.4	36.3	36.5	36.2

(200,200)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	78.4	78.3	78.4	78.4	78.4	78.4
10^{-3}	78.3	78.3	78.3	78.3	78.3	78.2
0.01	78.3	78.3	78.3	78.3	78.2	78.3
0.1	77.7	77.7	77.7	77.7	77.6	77.7
1	0.1	2.5	4.2	7.8	13.5	20.9
10	50.2	49.8	49.7	49.6	48.8	47.9

(200,100)

δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	81	81	81	80.6	80.9	80.9
10^{-3}	81	80.9	80.9	80.9	80.9	81
0.01	81.1	81	81	81	81	81.1
0.1	80.3	80.5	80.4	80.5	80.4	80.2
1	1.5	2.8	3.9	5.6	8.4	12.8
10	59.7	59.3	59.1	59	59.1	58.9

Figure 3.6 – Pourcentages de classes vides, pour des données simulées avec des proportions non équilibrées, obtenues par l'algorithme *V-Bayes* initialisés 10 fois avec $(H, L) = (4, 5)$ classes pour 100 matrices pour différentes valeurs de a (en ligne) et pour différentes tailles d'échantillons (en colonne).

Les résultats pour les cas de proportions identiques (figure 3.9) et différentes (figure 3.10) sont sensiblement les mêmes.

Prendre $a = 4$ a clairement un effet bénéfique contre la dégénérescence des classes, ce qui est en accord avec les recommandations prodiguées par Keribin et al. (2015).

3.2.3.b Choix empirique des hyperparamètres δ et ζ

Pour choisir les hyperparamètres δ et ζ, nous effectuons un plan d'expérience similaire où cette fois-ci nous fixons $a = 4$ et nous faisons varier ε varie entre 0 et 3 pour obtenir trois scénarios (voir figure 3.7) dans lesquelles les matrices sont plus ou moins simples à classifier via

$$\gamma = \begin{pmatrix} 15-2\varepsilon & 10-\varepsilon & 5 & 5 & 5 \\ 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon & 5 & 5 \\ 5 & 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon & 5 \\ 5 & 5 & 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon \end{pmatrix}.$$

Plus ε se rapproche de 5, moins les classes sont distinctes. Cette procédure est inspirée du protocole proposé par Lomet (2012) et repris par Keribin et al. (2015).

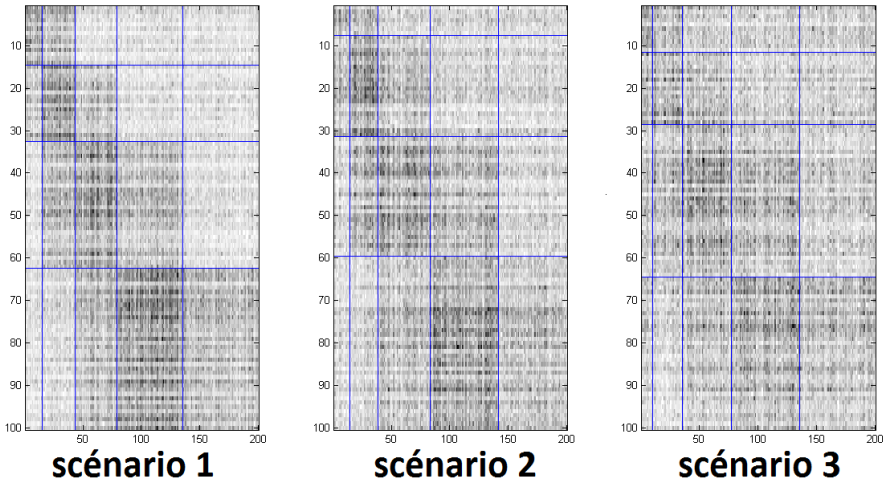


Figure 3.7 – *Exemple de matrices simulées c de données de comptage de tailles* $(100, 200)$ *et réorganisées pour les trois scénarios considérés.*

De même, nous utilisons l'algorithme *V-Bayes* pour des paramètres δ et ζ variant dans l'ensemble $(10^{-10}, 0.001, 0.01, 0.1, 1, 10)$. Pour chaque configuration, nous simulons 100 matrices. Comme l'algorithme *V-Bayes* est sensible aux initialisations, pour chaque matrice nous générons aléatoirement 10 initialisations et nous lançons avec chacune l'algorithme *V-Bayes*.

Dans les tableaux ci-dessous, nous recensons de manière analogue, le pourcentage de fois que l'algorithme *V-Bayes* a renvoyé une solution avec au moins une classe vide en ligne ou en colonne.

Les résultats pour les cas de proportions identiques (figure 3.9) et différentes (figure 3.10) sont sensiblement les mêmes.

Prendre $\delta = 1$ a clairement un effet bénéfique contre la dégénérescence des classes.

Ensuite, tout paramètre $\zeta \leqslant 1$ semble convenir également. Au vu de ces résultats, et afin d'effectuer un compromis entre un choix de loi a priori la moins informative possible et les expérimentations numériques, nous proposons de choisir les hyperparamètres $\delta = 1, \zeta = 0.01$ (voir figure 3.8).

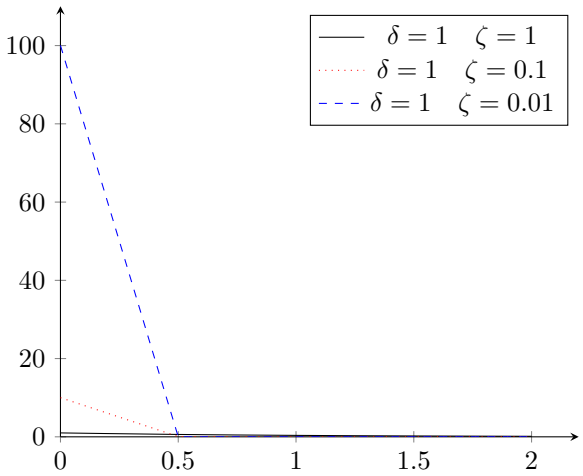


Figure 3.8 – Différentes densités de la distribution $\Gamma(\delta, \zeta)$ selon les valeurs de δ et ζ

proportions équilibrées

$(100, 200)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	6.9	6.9	7.1	8.3	13.1	36.6
10^{-3}	6.9	6.8	7.1	8.3	13.1	36.6
0.01	6.9	6.8	7.1	8.3	13.1	36.6
0.1	6.7	6.7	6.8	8.2	13.1	36.6
1	0	2.3	3.7	6.1	12.7	36.6
10	16.6	15.1	14.7	13.2	5.7	36.1

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	0.7	0.7	0.7	0.7	2.7	31.8
10^{-3}	0.7	0.7	0.7	0.7	2.7	31.8
0.01	0.7	0.7	0.7	0.7	2.7	31.8
0.1	0.5	0.5	0.6	0.8	2.6	31.7
1	0	0	0.1	0.6	2.3	31.4
10	5.7	5.3	4.9	3.4	0.2	28.2

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	35.7	35.7	35.7	35.7	35.7	35.7
10^{-3}	35.7	35.7	35.7	35.7	35.7	35.7
0.01	35.7	35.7	35.7	35.7	35.7	35.7
0.1	35.7	35.7	35.7	35.7	35.7	35.7
1	35.5	35.5	35.5	35.5	35.5	35.5
10	35.9	35.9	35.8	35.8	35.7	35.7

$(150, 150)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	10.3	10.4	10.7	11.5	17.5	42.3
10^{-3}	10.3	10.3	10.7	11.6	17.5	42.3
0.01	10.3	10.4	10.7	11.4	17.5	42.3
0.1	9.9	9.9	10.3	11.3	17.5	42.3
1	0	2	4	9.2	16.3	42.2
10	19.5	18.2	17.2	16.4	9.7	41.9

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	0.1	0.1	0.3	0.8	3.4	31.5
10^{-3}	0.1	0.1	0.3	0.8	3.4	31.5
0.01	0.1	0.1	0.1	0.7	3.4	31.3
0.1	0	0	0.1	0.6	3.4	31.3
1	0	0	0	0	2.9	31.4
10	8.3	7	6.4	4.3	0.4	29.1

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	37.5	37.5	37.5	37.5	37.5	37.5
10^{-3}	37.5	37.5	37.5	37.5	37.5	37.5
0.01	37.4	37.4	37.4	37.4	37.4	37.4
0.1	37.2	37.2	37.2	37.2	37.2	37.2
1	36.9	36.9	36.9	36.9	36.9	36.9
10	37.6	37.5	37.5	37.5	37.5	37.5

$(200, 100)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	78.4	78.3	78.4	78.4	78.4	78.4
10^{-3}	78.3	78.3	78.3	78.3	78.3	78.2
0.01	78.3	78.3	78.3	78.3	78.2	78.3
0.1	77.7	77.7	77.7	77.7	77.6	77.7
1	0.1	2.5	4.2	7.8	13.5	20.9
10	50.2	49.8	49.7	49.6	48.8	47.9

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	37	37	37	37.1	37.1	37
10^{-3}	37.1	37	37	37	37	37
0.01	37.1	37	37	37	37.1	37
0.1	36.8	36.7	36.7	36.8	36.8	36.8
1	34.7	34.6	34.6	34.6	34.6	34.6
10	35.7	35.5	35.5	35.5	35.4	35.3

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	37	37	37	37.1	37.1	37
10^{-3}	37.1	37	37	37	37	37
0.01	37.1	37	37	37	37.1	37
0.1	36.8	36.7	36.7	36.8	36.8	36.8
1	34.7	34.6	34.6	34.6	34.6	34.6
10	35.7	35.5	35.5	35.5	35.4	35.3

Figure 3.9 – Pourcentages de classes vides, pour des données simulées avec des proportions équilibrées, obtenues par l'algorithme *V-Bayes* initialisés 10 fois avec $(H, L) = (4, 5)$ classes pour 100 matrices pour les trois scénarios (en colonne) et pour différentes tailles d'échantillons (en ligne).

proportions non équilibrées

$(100, 200)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	59.5	59.5	59.6	59.5	59.5	59.5
10^{-3}	59.5	59.5	59.4	59.5	59.6	59.5
0.01	59.3	59.3	59.3	59.3	59.3	59.3
0.1	59	59	59	59	59	59.1
1	1.8	2.1	2.3	3.4	5.7	9.1
10	42.3	41.9	41.8	41.6	41.6	41.4

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	22.7	22.7	22.6	22.6	22.7	22.7
10^{-3}	22.7	22.6	22.6	22.6	22.6	22.6
0.01	22.5	22.6	22.6	22.5	22.5	22.5
0.1	22	22.1	22	22	22	22
1	13.9	13.9	13.8	13.8	13.8	13.9
10	20.7	20.4	20.3	20.3	20.3	20.3

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	58.2	58.2	58.2	58.2	58.2	58.2
10^{-3}	58.2	58.2	58.2	58.2	58.3	58.2
0.01	58.2	58.2	58.3	58.2	58.2	58.2
0.1	58.2	58.2	58.2	58.2	58.3	58.3
1	58.2	58.1	58.1	58.1	58.1	58.1
10	58.5	58.4	58.4	58.4	58.4	58.3

$(150, 150)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	69.9	69.9	69.9	69.9	69.9	69.9
10^{-3}	69.9	69.9	69.9	69.9	69.9	69.9
0.01	69.9	69.9	69.9	69.9	69.9	69.9
0.1	69	69	69	69	69	69
1	2	2.1	2.3	2.6	3.4	6
10	52.5	52.1	52.1	51.9	51.8	51.6

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	24.4	24.4	24.5	24.4	24.4	24.4
10^{-3}	24.4	24.4	24.4	24.4	24.4	24.4
0.01	24.4	24.3	24.3	24.3	24.3	24.3
0.1	23.6	23.6	23.6	23.6	23.6	23.6
1	15.6	15.6	15.6	15.6	15.6	15.6
10	28.4	27.7	27.6	27.3	27.1	27

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	56.9	56.9	56.9	56.9	56.9	56.9
10^{-3}	56.9	56.9	56.9	56.9	56.9	56.9
0.01	56.9	56.9	56.9	56.9	56.9	56.9
0.1	56.9	56.9	56.9	56.9	56.9	56.9
1	56.9	57	56.9	57	56.9	57
10	57.5	57.5	57.5	57.5	57.5	57.5

$(200, 100)$

scénario 1						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	77.2	77	77.2	77.2	77.2	77.4
10^{-3}	77	77.1	77.1	77.1	77.1	77.2
0.01	76.9	76.9	76.8	76.8	76.9	76.9
0.1	76.1	76.1	76.1	76.1	76.1	76.2
1	1.1	1.4	1.7	2.4	4.8	9
10	55.1	54.5	54.3	54.3	53.7	52.8

scénario 2						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	28.6	28.5	28.4	28.5	28.5	28.4
10^{-3}	28.5	28.4	28.4	28.4	28.4	28.4
0.01	28.4	28.5	28.4	28.4	28.4	28.4
0.1	27.5	27.5	27.6	27.5	27.5	27.6
1	13.3	13.3	13.3	13.3	13.4	13.3
10	28.2	27.2	27	27	26.8	26.6

scénario 3						
δ \ ζ	10^{-10}	10^{-3}	0.01	0.1	1	10
10^{-10}	56.9	56.9	56.9	56.9	56.9	56.9
10^{-3}	56.7	57.1	57	56.9	56.9	56.9
0.01	56.9	56.9	56.9	56.9	56.9	56.9
0.1	57	57	57	56.9	56.9	57
1	56.6	56.7	56.7	56.7	56.7	56.7
10	57.8	57.8	57.8	57.8	57.8	57.8

Figure 3.10 – Pourcentages de classes vides, pour des données simulées avec des proportions non équilibrées, obtenues par l'algorithme *V-Bayes* initialisés 10 fois avec $(H, L) = (4, 5)$ classes pour 100 matrices pour les trois scénarios (en colonne) et pour différentes tailles d'échantillons (en ligne).

3.3 Estimation des partitions et évaluation de leur qualité

3.3.1 Règle du Maximum A Posteriori (MAP)

Les probabilités conditionnelles $r_{jh} = \mathbb{P}(V_{jh} = 1|c;\theta)$ et $t_{k\ell} = \mathbb{P}(W_{k\ell} = 1|c;\theta)$ permettent d'attribuer une classe d'appartenance en ligne et en colonne à l'observation c_{jk} définies par les partitions v et w, à l'aide de la règle du Maximum A Posteriori (MAP) :

$$v_{jh} = \begin{cases} 1 & \text{si } \arg\max_{h'} r_{jh'} = h \\ 0 & \text{sinon,} \end{cases} \quad \text{et } w_{k\ell} = \begin{cases} 1 & \text{si } \arg\max_{\ell'} t_{k\ell'} = \ell \\ 0 & \text{sinon,} \end{cases}$$

En pratique, ces probabilités conditionnelles ne sont pas connues et sont remplacés par leurs estimations fournis par l'algorithme utilisé :

$$\widehat{v}_{jh} = \begin{cases} 1 & \text{si } \arg\max_{h'} \widehat{r}_{jh'} = h \\ 0 & \text{sinon,} \end{cases} \quad \text{et } \widehat{w}_{k\ell} = \begin{cases} 1 & \text{si } \arg\max_{\ell'} \widehat{t}_{k\ell'} = \ell \\ 0 & \text{sinon.} \end{cases}$$

3.3.2 Mesure d'information mutuelle généralisée pour coclustering proposée par Wyse et al. (2016)

Wyse et al. (2016) propose d'étendre la mesure d'information mutuelle généralisée introduite par Vinh et al. (2010) pour comparer deux partitions de coclustering. À l'origine, l'information mutuelle généralisée entre deux partitions $z = (z_1, \ldots, z_H)$ et $z' = (z'_1, \ldots, z'_{H'})$ d'un même ensemble $A = \{O_1, \ldots, O_I\}$ s'écrit :

$$\mathcal{I}(z, z') = \sum_{h,h'} P_{h,h'} \log\left(\frac{P_{h,h'}}{P_h P_{h'}}\right),$$

où

$$P_{h,h'} = \frac{1}{I}\sum_{i,i'}^{I} \mathbb{1}_{\{z_i=h, z'_{i'}=h'\}}, \quad P_h = \frac{1}{I}\sum_{i}^{I} \mathbb{1}_{\{z_i=h\}} \quad \text{et} \quad P_{h'} = \frac{1}{I}\sum_{i'}^{I} \mathbb{1}_{\{z'_{i'}=h'\}}.$$

Lorsque les deux partitions ne présentent pas le même nombre de classes, cette quantité est normalisée de la façon suivante :

$$\frac{\mathcal{I}(z, z')}{\max(\mathcal{H}(z), \mathcal{H}(z')},$$

où

$$\mathcal{H}(z) = -\sum_{h} P_h \log P_h, \text{ et } \mathcal{H}(z') = -\sum_{h'} P_{h'} \log P_{h'}.$$

Ainsi, la mesure proposée pour comparer deux partitions de coclustering ($z = (z_1, \ldots, z_H)$, $w = (w_1, \ldots, w_L)$) et ($z' = (z'_1, \ldots, z'_{H'})$, $w' = (w'_1, \ldots, w'_{L'})$) sur un ensemble $A \times B$ est une combinaison linéaire des mesures de comparaison entre les partitions z et z' et les partitions w et w' :

$$\mathcal{I}((z, w), (z', w')) = \mathcal{I}(z, z') + \mathcal{I}(w, w').$$

La valeur maximale est alors de 2 lorsque les partitions coïncident parfaitement à une permutation près et vaut 0 lorsque la correspondance entre les partitions est très faible.

En étendant de cette manière, des indices de classification simple à la classification croisée, la structure de coclustering du problème n'est pas préservée. En revanche, les deux indices que nous détaillons ensuite sont pensés d'un point de vue de coclustering et une comparaison entre ces deux façons de faire est présentée dans la section 3.3.4.c.

3.3.3 Erreur de classification croisée proposée par Lomet (2012)

L'erreur de classification étudiée dans la thèse de Lomet (Lomet (2012)) étudie le taux de mauvais classement des observations dans les blocs :

$$\mathrm{dist}_{(J,H)\times(K,L)}((v,w);(v',w')) = \min_{\sigma\in\mathfrak{S}(\{1,..H\})}\min_{\tau\in\mathfrak{S}(\{1,..L\})}(1-\tfrac{1}{J\times K}\sum_{j,h,k,\ell} v_{jh}v'_{j\sigma(h)}w_{k\ell}w'_{k\tau(\ell)}), \quad (3.3)$$

où $\mathfrak{S}(\{1,..H\})$ représente l'ensemble des permutations possibles de l'ensemble $\{1,\ldots,H\}$ permettant d'éviter le label switching et (v,w) et (v',w') des partitions de coclustering. Lomet (2012) montre que cette distance peut s'exprimer en fonction des distances simples relatives aux partitions en ligne et en colonne.

En effet, nous définissons de même la distance de classification sur les lignes entre deux matrices de partition

$$\mathrm{dist}_{J,H}(v;v') = 1 - \max_{\sigma\in\mathfrak{S}(\{1,..H\})}\frac{1}{J}\sum_{j,h} v_{jh}v'_{j\sigma(h)}, \quad (3.4)$$

Dans le cas où deux partitions à comparer ne comptent pas le même nombre de classes, nous prenons pour H le nombre maximal de classes et les classes supplémentaires sont supposées vides.

De manière symétrique, nous définissons la distance de classification sur les colonnes notée $\mathrm{dist}_{K,L}$.

Enfin, dans le cadre de la classification croisée, la distance de classification entre deux couples de partitions peut s'exprimer ainsi (voir figure 3.11) :

$$\mathrm{dist}_{(J,H)\times(K,L)}((v,w);(v',w')) = \mathrm{dist}_{J,H}(v;v') + \mathrm{d}_{K,L}(w;w') - \mathrm{dist}_{J,H}(v;v')\times\mathrm{dist}_{K,L}(w;w').$$

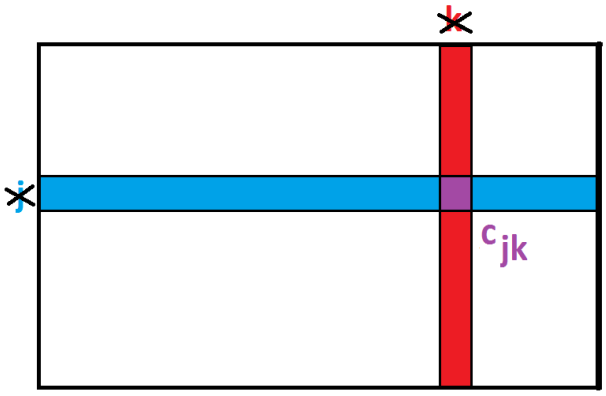


Figure 3.11 – Schéma explicatif de la distance de classification pour un tableau c.

Nous parlons alors d'*erreur de classification* lorsque cette fonction de coût mesure la différence entre le couple de partitions de référence (v^*,w^*) et une estimation $(\widehat{v},\widehat{w})$:

$$\mathrm{e}_{(J,H)\times(K,L)}((\widehat{v},\widehat{w}),(v^*,w^*)) \quad = \quad \mathrm{dist}_{(J,H)\times(K,L)}((\widehat{v},\widehat{w});(v^*,w^*)).$$

L'erreur de classification est comprise entre 0 et 1. Ainsi, l'observation c_{jk} n'est pas dans le bloc (h, ℓ) si la ligne j n'est pas dans la classe h *ou* si la colonne k n'est pas dans la classe ℓ. Se tromper dans la classification d'une colonne par exemple, revient à pénaliser toutes les cases de cette dernière, et d'augmenter l'erreur de $\frac{1}{K}$. Dans le cas où la ligne j (en bleu) et la colonne k (en rouge) sont mal classées, l'observation c_{jk} (en mauve) est comptée deux fois (voir figure 3.11), d'où la présence du terme de soustraction du produit des deux erreurs dans l'équation 3.5.

Remarque. Le cardinal de l'ensemble $\mathfrak{S}(\{1, ..H\})$ étant $H!$, le calcul numérique de cette distance devient difficile lorsque H est plus grand que 8.

3.3.4 Extension proposée : l'indice *Coclustering Adjusted Rand Index*

3.3.4.a Rand Index et Adjusted Rand Index

Rand (1971) a développé une mesure pour comparer des partitions entre elles. Plus précisément, soient deux partitions $z = (z_1, \ldots, z_H)$ et $z' = (z'_1, \ldots, z'_{H'})$ d'un même ensemble $A = \{O_1, \ldots, O_I\}$. z représente une partition de référence par exemple et z' un résultat d'une méthode de classification. Le *Rand index* s'écrit alors :

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{I}{2}}, \tag{3.5}$$

où,

- a représente le nombre de paires d'éléments qui sont placés dans la même classe dans z et dans z',
- b représente le nombre de paires d'éléments dans la même classe dans z mais pas dans la même classe dans z',
- c représente le nombre de paires d'éléments dans la même classe dans z' mais pas dans la même classe dans z,
- d le nombre de paires d'éléments qui sont dans différentes classes dans z et dans z'. Les variables a et d peuvent être interprétés comme des variables d'accord, et b et c comme des variables de désaccords.

Un tableau de contingence peut être introduit également. Soit $n^{zz'} = (n^{zz'}_{h,h'})_{H \times H'}$ une matrice où $n^{zz'}_{h,h'}$ représente le nombre d'éléments de A qui appartiennent à la fois à la classe z_h et à la classe $z'_{h'}$. Les marges lignes et colonnes $n^{zz'}_{h,\bullet}$ et $n^{zz'}_{\bullet,h'}$ représentent respectivement le nombre d'éléments dans la classe z_h et $z'_{h'}$. Nous avons la correspondance suivante :

- $a = \sum_{h}^{H} \sum_{h'}^{H'} \binom{n^{zz'}_{h,h'}}{2} = \frac{\sum_{h}^{H} \sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2 - I}{2}$,
- $b = \sum_{h}^{H} \binom{n^{zz'}_{h,\bullet}}{2} - a = \frac{\sum_{h}^{H} (n^{zz'}_{h,\bullet})^2 - \sum_{h}^{H} \sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2}{2}$,
- $c = \sum_{h'}^{H'} \binom{n^{zz'}_{\bullet,h'}}{2} - a = \frac{\sum_{h'}^{H'} (n^{zz'}_{\bullet,h'})^2 - \sum_{h}^{H} \sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2}{2}$,

- $d = \binom{I}{2} - a - b - c = \sum_{h}^{H} \binom{n_{h,\cdot}^{zz'}}{2} - \sum_{h'}^{H'} \binom{n_{\cdot,h'}^{zz'}}{2} + a = \frac{\sum_{h}^{H}\sum_{h'}^{H'} \left(n_{h,h'}^{zz'}\right)^2 + I^2 - \sum_{h}^{H} \left(n_{h,\cdot}^{zz'}\right)^2 - \sum_{h'}^{H'} \left(n_{\cdot,h'}^{zz'}\right)^2}{2}$.

Cet indice symétrique est compris entre 0 et 1 et prend la valeur 1 quand les deux partitions sont égales à une permutation près. En comparant des paires d'éléments, cet indice ne nécessite pas de passer en revue toutes les permutations des partitions étudiées, et son calcul est alors facile.

Cependant, la valeur attendue du *Rand index* pour deux partitions aléatoires ne prend pas une valeur constante, et de plus il est souvent concentré dans un petit intervalle proche de 1. L'*Adjusted Rand Index* (ARI) proposé par Hubert and Arabie (1985) permet de surmonter des inconvénients.Cette version corrigée suppose comme modélisation de l'aléatoire, la distribution hypergéométrique généralisée, soit que les partitions sont choisies aléatoirement en fixant le nombre d'éléments dans les classes.

La forme générale de cet indice est la suivante :

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}.$$

Cet indice est donc borné par 1, et prend cette valeur lorsque les deux permutations sont égales à une permutation près. Cet indice peut également prendre des valeurs négatives, ce qui traduit une très faible correspondance entre les deux partitions.

À partir de cette formule, Hubert and Arabie (1985) montre que l'indice peut s'écrire à l'aide du tableau de contingence :

$$ARI(z,z') = \frac{\sum_{h,h'} \binom{n_{h,h'}^{zz'}}{2} - \sum_{h} \binom{n_{h,\cdot}^{zz'}}{2} \sum_{h'} \binom{n_{\cdot,h'}^{zz'}}{2} / \binom{I}{2}}{\frac{1}{2}\left[\sum_{h} \binom{n_{h,\cdot}^{zz'}}{2} + \sum_{h'} \binom{n_{\cdot,h'}^{zz'}}{2}\right] - \left[\sum_{h} \binom{n_{h,\cdot}^{zz'}}{2} \sum_{h'} \binom{n_{\cdot,h'}^{zz'}}{2}\right] / \binom{I}{2}} \tag{3.6}$$

$$= \frac{2(ad-bc)}{b^2 + c^2 + 2ad + (a+d)(b+c)}. \tag{3.7}$$

3.3.4.b The Coclustering Adjusted Index

Nous proposons d'étendre cette formule d'un point de vue classification croisée. Ainsi, nous souhaitons comparer des paires de partitions qui définissent des blocs, et non plus des classes.

Définition 3.3.4.1. Soient $z = (z_1, \ldots, z_H)$ et $z' = (z'_1, \ldots, z'_{H'})$ deux partitions d'un ensemble $A = \{L_1, \ldots, L_I\}$ et $w = (w_1, \ldots, w_L)$ et $w' = (w'_1, \ldots, w'_{L'})$ deux partitions d'un ensemble $B = \{C_1, \ldots, C_J\}$. Notons qu'une observation de $A \times B$ est notée $x_{ij}, i = 1, \ldots, I; j = 1, \ldots, J$. Nous définissons un tableau de contingence $\tilde{n} = (\tilde{n}_{p,q})_{(H\times L)\times(H'\times L')}$ tel que $\tilde{n}_{p,q}$ représente le nombre d'éléments de l'ensemble $A \times B$ qui appartiennent à la fois au bloc $p = (h, \ell)$ et au bloc $q = (h', \ell')$. Les marges lignes et colonnes $\tilde{n}_{p,\cdot}$ et $\tilde{n}_{\cdot,q}$ représentent respectivement le nombre d'éléments dans le bloc p et le bloc q.

Il existe une bijection entre les indices i et les couples (h, ℓ) donnée par la relation $(p-1) = (h-1) \times L + (\ell - 1)$, de même pour les indices q et les couples (h', ℓ').

Remarquons que nous avons choisi la correspondance suivante entre l'indice p des lignes du tableau de contingence, et le bloc (h, ℓ) défini par (z, w) :

$$
\begin{array}{rcl}
1 & \leftrightarrow & (1,1) \\
2 & \leftrightarrow & (1,2) \\
3 & \leftrightarrow & (1,3) \\
\vdots & & \vdots \\
L & \leftrightarrow & (1,L) \\
L+1 & \leftrightarrow & (2,1) \\
\vdots & & \vdots \\
HL & \leftrightarrow & (H,L)
\end{array}
$$

Ainsi, grâce à ceci, le tableau de contingence peut être vu comme une matrice par blocs avec $H \times H'$ blocs de taille $L \times L'$ (voir tableau 3.1).

Bloc p\ Bloc q	$q=1$	$q=2$	...	$q=L'$	$q=L'+1$	...	$q=H'\times L'$	Marge
$p=1$	$\tilde{n}_{11}^{zwz'w'}$	$\tilde{n}_{12}^{zwz'w'}$	...	$\tilde{n}_{1,L'}^{zwz'w'}$	$\tilde{n}_{1,L'+1}^{zwz'w'}$	...	$\tilde{n}_{1H'\times L'}^{zwz'w'}$	$\tilde{n}_{1\bullet}^{zwz'w'}$
$p=2$	$\tilde{n}_{21}^{zwz'w'}$	$\tilde{n}_{22}^{zwz'w'}$	...	$\tilde{n}_{2,L'}^{zwz'w'}$	$\tilde{n}_{2,L'+1}^{zwz'w'}$	...	$\tilde{n}_{2,H\times L'}^{zwz'w'}$	$\tilde{n}_{2\bullet}^{zwz'w'}$
...	...	...	...	...	...	...	...	...
$p=L$	$\tilde{n}_{L,1}^{zwz'w'}$	$\tilde{n}_{L,2}^{zwz'w'}$	...	$\tilde{n}_{L,L'}^{zwz'w'}$	$\tilde{n}_{L,L'+1}^{zwz'w'}$	...	$\tilde{n}_{L,H\times L'}^{zwz'w'}$	$\tilde{n}_{L,\bullet}^{zwz'w'}$
$p=L+1$	$\tilde{n}_{L+1,1}^{zwz'w'}$	$\tilde{n}_{L+1,2}^{zwz'w'}$	...	$\tilde{n}_{L+1,L'}^{zwz'w'}$	$\tilde{n}_{L+1,L'+1}^{zwz'w'}$	...	$\tilde{n}_{L+1,H\times L'}^{zwz'w'}$	$\tilde{n}_{L+1,\bullet}^{zwz'w'}$
...	...	...	...	...	...	...	...	...
$p=H\times L$	$\tilde{n}_{H\times L,1}^{zwz'w'}$	$\tilde{n}_{H\times L,2}^{zwz'w'}$	...	$\tilde{n}_{H\times L,L'}^{zwz'w'}$	$\tilde{n}_{H\times L,L'+1}^{zwz'w'}$	...	$\tilde{n}_{H\times L,H'\times L'}^{zwz'w'}$	$\tilde{n}_{H\times L,\bullet}^{zwz'w'}$
Marge	$\tilde{n}_{\bullet,1}^{zwz'w'}$	$\tilde{n}_{\bullet,2}^{zwz'w'}$	...	$\tilde{n}_{\bullet,L'}^{zwz'w'}$	$\tilde{n}_{\bullet,L'+1}^{zwz'w'}$	...	$\tilde{n}_{\bullet,H'\times L'}^{zwz'w'}$	$\tilde{n}_{\bullet,\bullet}^{zwz'w'} = I \times J$

Table 3.1 – Notation pour le tableau de contingence pour comparer deux couples de partitions.

Une correspondance similaire peut être explicitée entre l'indice q et le bloc (h', ℓ') défini par (z', w'). Ainsi la notation $(h_p\ \ell_p)$ et $(h'_q\ \ell'_q)$ pourra être utilisée. Nous pouvons alors exprimer le *Coclustering Adjusted Rand Index* en fonction de $\tilde{n}^{zwz'w'}$.

Définition 3.3.4.2. Soient z, w, z', w' et $\tilde{n}^{zwz'w'}$ définis comme dans la définition 3.3.4.1. Le *Coclustering Adjusted Rand Index* (CARI) s'écrit :

$$
\text{CARI}((z,w),(z',w')) = \frac{\sum_{p,q}\binom{\tilde{n}_{p,q}^{zwz'w'}}{2} - \sum_{p}\binom{\tilde{n}_{p,\bullet}^{zwz'w'}}{2}\sum_{q}\binom{\tilde{n}_{\bullet,q}^{zwz'w'}}{2}/\binom{I\times J}{2}}{\frac{1}{2}\left[\sum_{p}\binom{\tilde{n}_{p,\bullet}^{zwz'w'}}{2} + \sum_{q}\binom{\tilde{n}_{\bullet,q}^{zwz'w'}}{2}\right] - \left[\sum_{p}\binom{\tilde{n}_{p,\bullet}^{zwz'w'}}{2}\sum_{q}\binom{\tilde{n}_{\bullet,q}^{zwz'w'}}{2}\right]/\binom{I\times J}{2}}. \quad (3.8)
$$

Tout comme l'ARI, cet indice est symmétrique et prend la valeur 1 quand les couples de partitions sont égales à une permutation près. Contrairement à l'indice proposé par Lomet (2012), aucune convention n'est requise lorsque le nombre de classes est différent pour les partitions et son calcul ne nécessite pas de passer en revue toutes les permutations des

partitions et peut être calculé même si le nombre de classes dépasse neuf. Cependant, la complexité naïve pour calculer $\tilde{n}^{zwz'w'}$ reste substantielle et vaut $H \times L + H' \times L' + H \times L \times H' \times L'$. Ces deux indices seront expérimentés sur données simulées et données réelles dans la section 4.2.4. Heureusement, le théorème suivant nous permet de calculer $\tilde{n}^{zwz'w'}$ de manière plus rapide et rendre ainsi CARI plus compétitif :

Théorème 3.3.4.3. *Soient $z, w, z', w', \tilde{n}^{zwzw'}, n^{zz'}$, et $n^{ww'}$ définis comme dans la définition 3.3.4.1. Nous avons alors la relation suivante,*

$$n^{zwzw'} = n^{zz'} \otimes n^{ww'}, \tag{3.9}$$

où $\otimes$ représente le produit de Kronecker entre deux matrices.

Grâce à ce théorème, le tableau de contingence $\tilde{n}^{zwzw'}$ est maintenant défini selon l'équation 3.9. Son calcul est maintenant plus efficace . De plus, même si le produit de Kronecker n'est pas commutatif, il se comporte bien avec les opérateurs de transposée et de marge et les propriétés initiales de CARI sont conservées :

Corollaire 3.3.4.4. *1. $\forall (p,q) \in (H \times L) \times (H' \times L')$, Nous avons les relations suivantes, $\tilde{n}^{zwzw'}_{\bullet,q} = n^{zz'}_{\bullet h'_q} \otimes n^{ww'}_{\bullet,\ell'_q}$ and $\tilde{n}^{zwzw'}_{p,\bullet} = n^{zz'}_{h_p,\bullet} \otimes n^{ww'}_{\ell_p,\bullet}$.*

2. Le CARI associé avec le tableau de contingence $\tilde{n}^{zwzw'}$ défini par 3.9 reste symétrique, soit

$$CARI((z,w),(z',w')) = CARI((z',w'),(z,w)).$$

3.3.4.c Exemples

Exemple 1 : comparaison de couples de partitions égaux à une permutation près

Considérons les couples de partitions suivants $(z,w) = ((1,1,3,2),(1,2,1,4,3,))$ et $(z',w') = ((2,2,1,3),(2,1,2,3,4))$. Nous remarquons que les partitions z et z' sont égales à une permutation près, de même que w et w'. Le tableau de contingence (voir tableau 3.2) associé à CARI$((z,w),(z',w'))$ est donc de taille de $(3 \times 4, 3 \times 4)$.

Ainsi, le CARI$((z,w),(z',w'))$ vaut $\frac{11-121/190}{1/2\times 22-121/190} = 1$.

Bloc	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	Marge
$(1,1)$	0	0	0	0	0	4	0	0	0	0	0	0	4
$(1,2)$	0	0	0	0	2	0	0	0	0	0	0	0	2
$(1,3)$	0	0	0	0	0	0	0	2	0	0	0	0	2
$(1,4)$	0	0	0	0	0	0	2	0	0	0	0	0	2
$(2,1)$	0	0	0	0	0	0	0	0	0	2	0	0	2
$(2,2)$	0	0	0	0	0	0	0	0	1	0	0	0	1
$(2,3)$	0	0	0	0	0	0	0	0	0	0	0	1	1
$(2,4)$	0	0	0	0	0	0	0	0	0	0	1	0	1
$(3,1)$	0	2	0	0	0	0	0	0	0	0	0	0	2
$(3,2)$	1	0	0	0	0	0	0	0	0	0	0	0	1
$(3,3)$	0	0	0	1	0	0	0	0	0	0	0	0	1
$(3,4)$	0	0	1	0	0	0	0	0	0	0	0	0	1
Marge	1	2	1	1	2	4	2	2	1	2	1	1	20

Table 3.2 – Tableau de contingence $\tilde{n}$ associé au CARI$((z,w),(z',w'))$.

Exemple 2 : Comparaison de couples de partitions comportant un nombre de classes différent

Considérons les couples de partitions suivants $(z, w) = ((1, 1, 2, 2, 1), (1, 1, 2, 1, 1, 2))$ et $(z', w') = ((1, 1, 2, 1, 1), (1, 1, 2, 1, 3, 2))$. Remarquons que les partitions w et w' ne possèdent pas le même nombre de classes.

Les différentes tableaux de contingence relatifs à $\text{ARI}(z, z')$, $\text{ARI}(w, w')$ et $\text{CARI}((z, w), (z', w'))$ sont représentées respectivement par les figures 3.3 et 3.4. Nous constatons bien que

$$\tilde{n}^{zwz'w'} = n^{zz'} \otimes n^{ww'}.$$

La valeur des différents *ARI* est disponible dans le tableau 3.5. Nous remarquons que la valeur de $\text{CARI}((z, w), (z', w'))$ semble se rapprocher de la moyenne des valeurs des *ARI* simples.

Classe	1	2	Marge
1	3	0	3
2	1	1	2
Marge	4	1	5

Classe	1	2	3	Marge
1	3	0	1	4
2	0	2	0	2
Marge	3	2	1	6

Table 3.3 – Tableau de contingence n associé à l'$ARI(z, z')$ (à gauche) et à l'$ARI(w, w')$ (à droite).

Bloc	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$	Marge
$(1,1)$	9	0	3	0	0	0	12
$(1,2)$	0	6	0	0	0	0	6
$(2,1)$	3	0	1	3	0	1	8
$(2,2)$	0	2	0	0	2	0	4
Marge	12	8	4	3	2	1	30

Table 3.4 – Tableau de contingence $\tilde{n}$ associé à $\text{CARI}((z, w), (z', w'))$.

Valeur	$\text{ARI}(z, z')$	$\text{ARI}(w, w')$	$\text{CARI}((z, w), (z', w'))$
	0.2308	0.5872	0.4208

Table 3.5 – Comparaison des valeurs de $\text{ARI}(z, z')$, $\text{ARI}(w, w')$ et $\text{CARI}((z, w), (z', w'))$.

Exemple 3 : comparaison entre CARI$((z, w), (z', w'))$ et $\mathcal{I}((z, w), (z', w'))$

Considérons les couples de partitions suivants $(z, w) = ((1, 2, 2, 13), (1, 2, 2, 3, 3,))$ et $(z', w') = ((1, 2, 2, 1, 3), (1, 1, 2, 3, 2))$. Nous remarquons que les partitions z et z' sont les mêmes. Par conséquent, l'erreur proposée par Wyse et al. (2016) sera d'au minimum 1 alors que les partitions w et w' sont très peu concordantes. Nous avons ainsi

$$\text{CARI}((z, w), (z', w')) = 0.2910, \quad \text{et} \quad \mathcal{I}((z, w), (z', w')) = 1.4744.$$

L'indice $\mathcal{I}((z, w), (z', w'))$ qui vaut au maximum 2 indique qu'il existe une assez forte correspondance entre ces deux partitions de coclustering alors que l'indice CARI indique une

assez faible correspondance. Même si les partitions en lignes sont les mêmes, les partitions en colonne sont très discordantes. D'un point de vue de classification croisée, ces couples de partitions sont peu concordantes, ce que traduit l'indice CARI mais que n'arrive pas à retrouver l'indice $\mathcal{I}$ du fait de sa construction comme simple combinaison linéaire des indices de classification simple pour les lignes et les colonnes.

3.4 Annexes

3.4.1 Formulaire de l'échantillonneur de Gibbs

Dans cette section, nous allons détailler toutes les formules obtenues dans les étapes de l'échantillonneur de Gibbs. Nous avons une symétrie entre v et w et nous ne détaillerons que le cas de v :

$$\begin{aligned}
\mathbb{P}(v_j = h|w,c,\theta) &= \frac{\mathbb{P}(c_{j\bullet}|v_j = h,w,c,\theta)\mathbb{P}(v_j = h|\theta)}{\mathbb{P}(c_{j\bullet}|w,\theta)} \\
&= \frac{\rho_h \prod_k \mathbb{P}(c_{jk}|v_j = h,w,c,\theta)}{\sum_{h'} \rho_{h'} \prod_k \mathbb{P}(c_{jk}|v_j = h',w,c,\theta)} \\
&= \frac{\rho_h \prod_{k,\ell} \phi(c_{jk};\mu_j\nu_k\gamma_{h\ell})^{w_{k\ell}}}{\sum_{h'} \rho_{h'} \prod_{k,\ell} \phi(c_{jk};\mu_j\nu_k\gamma_{h'\ell})^{w_{k\ell}}} \\
&= \frac{\rho_h e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell} w_{k\ell}} \prod_\ell \left[(\gamma_{h\ell})^{\sum_k w_{k\ell} c_{jk}}\right] \cancel{(\mu_j)^{\sum_k w_{k\bullet} c_{jk}} \prod_k \nu_k^{w_{k\bullet} c_{jk}}}}{\sum_{h'} \rho_{h'} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h'\ell} w_{k\ell}} \prod_\ell \left[(\gamma_{h'\ell})^{\sum_k w_{k\ell} c_{jk}}\right] \cancel{(\mu_j)^{\sum_k w_{k\bullet} c_{jk}} \prod_k \nu_k^{w_{k\bullet} c_{jk}}}}
\end{aligned}$$

Par symétrie :

$$\mathbb{P}(w_j = \ell|v,c,\theta) = \frac{\tau_\ell e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell} v_{jh}} \prod_h (\gamma_{h\ell})^{\sum_j v_{jh} c_{jk}}}{\sum_{\ell'} \tau_{\ell'} e^{-\nu_k \sum_j \mu_j \sum_h \gamma_{h\ell'} v_{jh}} \prod_h (\gamma_{h\ell'})^{\sum_j v_{jh} c_{jh}}}$$

De même, il y a une symétrie entre ρ et τ. Nous avons :

$$\begin{aligned}
\mathbb{P}(\rho|v,w,\tau,\gamma_{h\ell},c) &\propto \mathbb{P}(v|\rho,w,\tau,\gamma_{h\ell},c)p(\rho|w,\tau,\gamma_{h\ell},c) \\
&\propto \mathbb{P}(v|\rho)p(\rho) \\
&\propto \prod_{j,h} \rho_h^{v_{jh}} \prod_h \rho_h^{a-1} \\
&\propto \prod_h \rho_h^{\left(a+\sum_j v_{jh}\right)-1} \\
&\propto \prod_h \rho_h^{(a+v_{\bullet h})-1}
\end{aligned}$$

en notant $v_{\bullet h} = \sum_j v_{jh}$. Donc, nous avons :

$$\rho|v,w,\tau,\gamma_{h\ell},c \sim \mathcal{D}\left(v_{\bullet 1} + a,\ldots,v_{\bullet H} + a\right)$$

De même, en notant $w_{\bullet\ell} = \sum_k w_{k\ell}$:

$$\tau|v, w, \rho, \gamma_{h\ell}, c \sim \mathcal{D}\left(w_{\bullet 1} + a, \ldots, w_{\bullet L} + a\right)$$

Remarque. En pratique, pour simuler une loi de Dirichlet, nous utilisons la propriété suivante : si pour $i \in \{1, \cdots, n\}$, $X_i \sim \Gamma(\delta_i, 1)$, alors $(\frac{X_1}{\sum_i X_i}, \cdots, \frac{X_n}{\sum_i X_i}) \sim \mathcal{D}(\delta_1, \cdots, \delta_n)$.

Il reste à calculer les probabilités a posteriori de $(\gamma_{h\ell})$:

$$\begin{aligned}
\mathbb{P}(\gamma|v, w, c) &\propto \mathbb{P}(c|v, w, \gamma)p(\gamma) \\
&\propto \prod_{j,k,h,\ell} \mathbb{P}(c_{jk}|\gamma_{h\ell})^{v_{jh}w_{k\ell}} \prod_{h,\ell} p(\gamma_{h\ell}) \\
&\propto \prod_{j,k,h,\ell} \left(e^{-\mu_j \nu_k \gamma_{h\ell}} \frac{(\mu_j \nu_k \gamma_{h\ell})^{c_{jk}}}{c_{jk}!} \right)^{v_{jh}w_{k\ell}} \times \prod_{h,\ell} \left(\frac{\zeta^\delta}{\Gamma(\delta)} \gamma_{hl}^{\delta-1} e^{-\zeta\gamma_{h\ell}} \right) \\
&\propto \prod_{h,\ell} \left[\gamma_{h\ell}^{\left(\delta + \sum_{j,k} v_{jh}w_{k\ell}c_{jk}\right)-1} e^{-\left(\zeta + \sum_{j,k} v_{jh}w_{k\ell}\right)\gamma_{h\ell}} \right]
\end{aligned}$$

Nous voyons que les paramètres pour chaque bloc sont indépendants et nous avons :

$$\gamma_{h\ell}|v, w, c \sim \Gamma\left(\delta + \sum_{j,k} v_{jh}w_{k\ell}c_{jk}, \zeta + \sum_{j,k} v_{jh}w_{k\ell}\mu_j\nu_k\right).$$

3.4.2 Formulaire de l'algorithme *V-Bayes*

L'étape d'estimation consiste à calculer $r_{jh}^{(e+1)} = \mathbb{P}(v_{jh} = 1|c; \theta^{(d)})$ et $t_{k\ell}^{(e+1)} = \mathbb{P}(w_{k\ell} = 1|c; \theta^{(d)})$. Comme les calculs sont symétriques, nous ne regarderons que le premier cas. Nous reprenons le point de vue abordé dans Govaert and Nadif (2013) consistant à maximiser alternativement :

$$\begin{aligned}
g(r, t, \gamma, \lambda, \zeta) &= \mathbb{E}_{(V,W)\sim q_{vw}} \left[\log \left(\frac{p(c, V, W; \theta)}{q_{vw}(V, W)} \right) \right] + \sum_j \lambda_j \left(\sum_h r_{jh} - 1 \right) + \sum_k \zeta_k \left(\sum_\ell t_{k\ell} - 1 \right) + \log p(\theta) \\
&= \sum_{j,h} r_{jh} \log \rho_h + \sum_{k,\ell} t_{k\ell} \log \tau_\ell + \sum_{j,k,h,\ell} r_{jh} t_{k\ell} \log \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell}) \\
&\quad - \sum_{j,h} r_{jh} \log r_{jh} - \sum_{k,\ell} t_{k\ell} \log t_{k\ell} + \sum_j \lambda_j \left(\sum_h r_{jh} - 1 \right) + \sum_k \zeta_k \left(\sum_\ell t_{k\ell} - 1 \right) + \log p(\theta).
\end{aligned}$$

En dérivant, nous avons :

$$\begin{aligned}
\frac{\partial}{\partial r_{jh}} g(r, t, \gamma, \lambda, \zeta) = 0 &\Leftrightarrow \log \left(\rho_h \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell})^{t_{k\ell}} \right) - \log r_{jh} - 1 + \lambda_j = 0 \\
&\Leftrightarrow \log r_{jh} = \log \left(\rho_h \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell})^{t_{k\ell}} \right) - 1 + \lambda_j \\
&\Leftrightarrow r_{jh} = e^{\lambda_j - 1} \rho_h \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell})^{t_{k\ell}}
\end{aligned}$$

En dérivant sur λ, on obtient que $\sum_h r_{jh} = 1$ ce qui donne finalement :

$$r_{jh} = \frac{\rho_h \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell})^{t_{k\ell}}}{\sum_{h'} \rho_{h'} \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h'\ell})^{t_{k\ell}}}.$$

En remplaçant ϕ par son expression, nous obtenons par suite,

$$\begin{aligned}
r_{jh} &= \frac{\rho_h \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{k\ell})^{t_{k\ell}}}{\sum_{h'} \rho_{h'} \prod_{k,\ell} \phi(c_{jk}; \mu_j \nu_k \gamma_{h'\ell})^{t_{k\ell}}} \\
&= \frac{\rho_h \prod_{k,\ell} \left(e^{-\mu_j \nu_k \gamma_{h\ell}} \frac{(\mu_j \nu_k \gamma_{h\ell})^{c_{jk}}}{c_{jk}!} \right)^{t_{k\ell}}}{\sum_{h'} \rho_{h'} \prod_{k,\ell} \left(e^{-\mu_j \nu_k \gamma_{h'\ell}} \frac{(\mu_j \nu_k \gamma_{h'\ell})^{c_{jk}}}{c_{jk}!} \right)^{t_{k\ell}}} \\
&= \frac{\cancel{\prod_{k,\ell} \left[\left(\frac{1}{c_{jk}!} \right)^{t_{k\ell}} \mu_j^{t_{k\ell} c_{jk}} \nu_k^{t_{k\ell} c_{jk}} \right]} \rho_h e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell} t_{k\ell}} \prod_\ell (\gamma_{h\ell})^{\sum_k t_{k\ell} c_{jk}}}{\cancel{\prod_{k,\ell} \left[\left(\frac{1}{c_{jk}!} \right)^{t_{k\ell}} \mu_j^{t_{k\ell} c_{jk}} \nu_k^{t_{k\ell} c_{jk}} \right]} \sum_{h'} \rho_{h'} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell} t_{k\ell}} \prod_\ell (\gamma_{h'\ell})^{\sum_k t_{k\ell} c_{jk}}} \\
&= \frac{\rho_h e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell} t_{k\ell}} \prod_\ell (\gamma_{h\ell})^{\sum_k t_{k\ell} c_{jk}}}{\sum_{h'} \rho_{h'} e^{-\mu_j \sum_k \nu_k \sum_\ell \gamma_{h\ell} t_{k\ell}} \prod_\ell (\gamma_{h'\ell})^{\sum_k t_{k\ell} c_{jk}}}.
\end{aligned}$$

Nous obtenons de même la formule de $t_{k\ell}$ présente dans le schéma de l'algorithme.

Pour l'étape de maximisation, nous allons chercher à maximiser une fonction similaire, en ajoutant cette fois-ci le lagrangien relatif à ρ et τ:

$$\begin{aligned}
f(\gamma_{h\ell}, \lambda, \zeta) &= \sum_{j,h} r_{jh} \log \rho_h + \sum_{j,\ell} t_{j\ell} \log \rho_\ell + \sum_{j,k,h,\ell} r_{jh} t_{k\ell} \log \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell}) - \sum_{j,h} r_{jh} \log r_{jh} - \sum_{k,\ell} t_{k\ell} \log t_{k\ell} \\
&\quad + \log p(\rho) + \log p(\tau) + \log p(\gamma_{h\ell}) + \lambda \left(\sum_h \rho_h - 1 \right) + \zeta \left(\sum_\ell \tau_\ell - 1 \right) \\
&= \sum_{j,h} r_{jh} \log \rho_h + \sum_{k,\ell} t_{k\ell} \log \rho_\ell + \sum_{j,k,h,\ell} r_{jh} t_{k\ell} \log \phi(c_{jk}; \mu_j \nu_k \gamma_{h\ell}) - \sum_{j,h} r_{jh} \log r_{jh} - \sum_{k,\ell} t_{k\ell} \log t_{k\ell} \\
&\quad + \log \frac{\Gamma(Ha)}{\Gamma(a)^H} + (a-1) \sum_h \log \rho_h + \log \frac{\Gamma(La)}{\Gamma(a)^L} + (a-1) \sum_\ell \log \tau_\ell \\
&\quad + \sum_{h,\ell} \log p(\gamma_{h\ell}) + \lambda \left(\sum_h \rho_h - 1 \right) + \zeta \left(\sum_\ell \tau_\ell - 1 \right)
\end{aligned}$$

La maximisation sur ρ et τ étant la même, nous ferons uniquement celle sur ρ_h :

$$\begin{aligned}
\frac{\partial}{\partial \rho_h} f(\gamma_{h\ell}, \lambda, \zeta) = 0 &\Leftrightarrow \frac{1}{\rho_h} \sum_j r_{jh} + \frac{a-1}{\rho_h} + \lambda = 0 \\
&\Leftrightarrow \frac{a - 1 + \sum_j r_{jh}}{\rho_h} = -\lambda \\
&\Leftrightarrow \rho_h = \frac{a - 1 + \sum_j r_{jh}}{-\lambda}.
\end{aligned}$$

En dérivant par rapport à λ, nous obtenons que $\sum_h \rho_h = 1$ et nous avons :

$$\rho_h = \frac{a - 1 + \sum_j r_{jh}}{H(a-1) + J}.$$

De même, nous avons :

$$\tau_\ell = \frac{a - 1 + \sum_k t_{k\ell}}{L(a-1) + K}.$$

Nous dérivons enfin par rapport au paramètre $\gamma_{h\ell}$:

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{h\ell}} h(\rho, \tau, \gamma_{h\ell}, \lambda, \mu) = 0 \quad &\Leftrightarrow \quad -\sum_{j,k} \mu_j \nu_k r_{jh} t_{k\ell} + \frac{1}{\gamma_{h\ell}} \sum_{j,k} r_{jh} t_{k\ell} c_{jk} + \frac{\delta - 1}{\gamma_{h\ell}} - \zeta = 0 \\
&\Leftrightarrow \quad \gamma_{h\ell} \left(\zeta + \sum_{j,k} \mu_j \nu_k r_{jh} t_{k\ell} \right) = \delta - 1 + \sum_{j,k} r_{jh} t_{k\ell} c_{jk} \\
&\Leftrightarrow \quad \gamma_{h\ell} = \frac{\delta - 1 + \sum_{j,k} r_{jh} t_{k\ell} c_{jk}}{\zeta + \sum_{j,k} \mu_j \nu_k r_{jh} t_{k\ell}}.
\end{aligned}$$

3.4.3 Preuve du théorème 3.3.4.3

Commençons par remarquer que nous avons aligné les indices de la façon suivante :

$$\begin{aligned}
1 &\leftrightarrow (1,1) \\
2 &\leftrightarrow (1,2) \\
3 &\leftrightarrow (1,3) \\
&\vdots \quad\quad \vdots \\
L &\leftrightarrow (1,L) \\
L+1 &\leftrightarrow (2,1) \\
&\vdots \quad\quad \vdots \\
HL &\leftrightarrow (H,L)
\end{aligned}$$

Ainsi, nous avons l'équivalence suivante :
Pour tout $p \in \{1, \ldots, HL\}$, le couple (h, ℓ) associé est le quotient+1 et le reste+1 de la division euclidienne de $(p-1)$ par L. Autrement dit, nous avons :

$$(p-1) = (h-1) \times L + (\ell - 1).$$

En effet, nous pouvons exhiber les relations suivantes :

$$\begin{array}{rcl}
1 & \to & (1-1) = 0 = 0 \times L + 0 \to (h,\ell) = (1,1) \\
2 & \to & (2-1) = 1 = 0 \times L + 1 \to (h,\ell) = (1,2) \\
3 & \to & (3-1) = 2 = 0 \times L + 2 \to (h,\ell) = (1,3) \\
\vdots & & \vdots \\
L & \to & (L-1) = 0 \times L + (L-1) \to (h,\ell) = (1,L) \\
L+1 & \to & (L+1-1) = L = 1 \times L + 0 \to (h,\ell) = (2,1) \\
\vdots & & \vdots \\
HL & \to & (HL-1) = (H-1) \times L + (L-1) \to (h,\ell) = (H,L)
\end{array}$$

Corollaire 3.4.3.1. *Il y a une bijection entre chaque indice p et les couples (h,ℓ).*

Preuve. Ceci est vrai par l'existence et l'unicité du couple obtenu dans la division euclidienne. □

Ensuite, remarquons la proposition suivante :

Proposition 3.4.3.2. *Nous avons pour tout couple d'indices p et q associés aux couples de classes (h,ℓ) et (h',ℓ') la formule suivante :*

$$\widetilde{n}_{p,q} = n_{h,h'}^{zz'} n_{\ell,\ell'}^{ww'}.$$

Preuve. Pour montrer ceci, nous commençons par remarquer que la case x_{ij} est dans le bloc (h,ℓ) si et seulement si la ligne i est dans la classe h et la colonne j est dans la classe ℓ.

Du coup, nous avons :

$$\begin{array}{rcl}
\{x_{ij} | x_{ij} \in bloc_{(h,\ell)} \cap bloc_{(h,\ell)}\} & = & \{x_{ij} | x_{ij} \in bloc_{(h,\ell)} \text{ et } x_{ij} \in bloc_{(h,\ell)}\} \\
& = & \{x_{ij} | i \in classe_h \text{ et } j \in classe_\ell \text{ et } i \in classe_{h'} \text{ et } j \in classe_{\ell'}\} \\
& = & \{x_{ij} | i \in classe_h \cap classe_{h'} \text{ et } j \in classe_\ell \cap classe_{\ell'}\}
\end{array}$$

Et nous obtenons le résultat.

□

Nous pouvons conclure en couplant les deux résultats précédents,

Preuve.

$$
\begin{aligned}
\widetilde{n} &= \begin{pmatrix}
\widetilde{n}_{1,1} & \widetilde{n}_{1,2} & \cdots & \widetilde{n}_{1,L'} & \widetilde{n}_{1,L'+1} & \cdots & \widetilde{n}_{1,H'L'} \\
\widetilde{n}_{2,1} & \widetilde{n}_{2,2} & \cdots & \widetilde{n}_{2,L'} & \widetilde{n}_{2,L'+1} & \cdots & \widetilde{n}_{2,H'L'} \\
\vdots & \vdots & \ddots & & & & \vdots \\
\widetilde{n}_{L,1} & \widetilde{n}_{L,2} & \cdots & \widetilde{n}_{L,L'} & \widetilde{n}_{L,L'+1} & \cdots & \widetilde{n}_{L,H'L'} \\
\widetilde{n}_{L+1,1} & \widetilde{n}_{L+1,2} & \cdots & \widetilde{n}_{L+1,L'} & \widetilde{n}_{L+1,L'+1} & \cdots & \widetilde{n}_{L+1,H'L'} \\
\vdots & \vdots & & & & \ddots & \vdots \\
\widetilde{n}_{HL,1} & \widetilde{n}_{HL,2} & \cdots & \widetilde{n}_{HL,L'} & \widetilde{n}_{HL,L'+1} & \cdots & \widetilde{n}_{HL,H'L'}
\end{pmatrix} \\
&= \begin{pmatrix}
n_{1,1}^{zz'}n_{1,1}^{ww'} & n_{1,1}^{zz'}n_{1,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{1,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{1,H'}^{ww'} \\
n_{1,1}^{z,z'}n_{2,1}^{ww'} & n_{1,1}^{zz'}n_{2,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{2,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{2,H'}^{ww'} \\
\vdots & \vdots & \ddots & & & & \vdots \\
n_{1,1}^{zz'}n_{H,1}^{ww'} & n_{1,1}^{zz'}n_{H,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{H,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{H,H'}^{ww'} \\
n_{2,1}^{zz'}n_{1,1}^{ww'} & n_{2,1}^{zz'}n_{1,2}^{ww'} & \cdots & n_{2,1}^{z,z'}n_{1,H'}^{ww'} & n_{2,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{2,L'}^{zz'}n_{1,H'}^{ww'} \\
\vdots & \vdots & & & & \ddots & \vdots \\
n_{L,1}^{zz'}n_{H,1}^{ww'} & n_{L,1}^{zz'}n_{H,2}^{ww'} & \cdots & n_{L,1}^{zz'}n_{HH'}^{ww'} & n_{L,2}^{zz'}n_{H,1}^{ww'} & \cdots & n_{L,L'}^{zz'}n_{H,H'}^{ww'}
\end{pmatrix} \\
&= \begin{pmatrix}
n_{1,1}^{zz'}n^{ww'} & n_{1,2}^{zz'}n^{ww'} & \cdots & n_{1,L'}^{zz'}n^{ww'} \\
n_{2,1}^{zz'}n^{ww'} & n_{2,2}^{zz'}n^{ww'} & \cdots & n_{2,L'}^{zz'}n^{ww'} \\
\vdots & & \ddots & \vdots \\
n_{L,1}^{zz'}n^{ww'} & n_{L,2}^{zz'}n^{ww'} & \cdots & n_{L,L'}^{zz'}n^{ww'}
\end{pmatrix} \\
&= n^{zz'} \otimes n^{ww'}.
\end{aligned}
$$

□

3.4.4 Preuve du corollaire 3.3.4.4

Preuve.

1. C'est une conséquence directe des propriétes connues du produit de Kronecker.
2. La preuve repose sur le lemme suivant :
   Lemma 3.1. Soient $z, w, z', w', n^{zz'}$, et $n^{ww'}$ définis comme dans la définition 1 et $\tilde{n}^{zwzw'}$ défini selon le théorème 3.3.1. Nous avons alors,
   $$n^{z'w'zw} = t(n^{zwz'w'}). \tag{3.10}$$
   Preuve. Grâce à la propriété du produit de Kronecker concernant l'opérateur transposée nous avons,
   $$
   \begin{aligned}
   n^{z'w'zw} &= t(n^{zz'}) \otimes t(n^{ww'}) \\
   &= t\left(n^{zz'} \otimes n^{ww'}\right) \\
   &= t(n^{zwz'w'}).
   \end{aligned}
   $$

□

Ainsi, si le CARI(z,w),(z',w') est calculé selon le théorème 3.3.4.3, la transposée de $n^{zz'}$ et de $n^{ww'}$ sont considérées pour générer $n^{z'w'zw}$. En définissant $n^{zwz'w'}$ selon le théorème 3.3.4.3, considérer (z,w),(z',w') ou (z',w'),(z,w), c'est considérer $n^{zwz'w'}$ ou sa transposée. Or, comme les marges jouent un rôle symétrique dans la définition de l'indice CARI (voir équation 3.8), en considérant $n^{zwz'w'}$ ou sa transposée, l'indice CARI reste inchangé.

□

4

Sélection de modèles pour le LBM Poisson normalisé : aspects théoriques et algorithmiques

L'estimation des paramètres dans le cadre du modèle des blocs latents a été présenté dans les chapitres précédents et le nombre de composantes en ligne et en colonne y était supposé connu. Maintenant, l'objectif est de concevoir des classes en ligne et en colonne lorsque les labels n'ont pas été observés, ce qui est généralement le cas. Ainsi, nous nous intéressons dans ce chapitre à l'évaluation de ce nombre de classes et nous allons développer le critère *Integrated Completed Likelihood* (ICL) dans le cadre du modèles étudié.

Dans le cadre des modèles de mélanges simples, Biernacki et al. (2000) proposent le critère ICL, une alternative au critère *Bayesian Information Criterion* noté BIC qui est consistant lorsque le vrai modèle est dans la collection de modèles considérés (Keribin (2000)) mais qui tend à surestimer le nombre de composantes lorsque le modèle qui a généré les données n'est pas dans la collection de modèles considérée. Ce critère se situe dans un cadre bayésien et l'objectif est de trouver le modèle $\mathcal{M}$ caractérisé par un nombre de composantes H

maximisant la vraisemblance intégrée complète :

$$(\widehat{H}, \widehat{\mathcal{M}}) \in \underset{\mathcal{M},H}{\operatorname{argmax}} \quad p(c, v^*|\mathcal{M}),$$

où c est la matrice d'observations et v^* la partition inhérente aux données.

En pratique, la partition v^* est inconnue et les auteurs proposent de l'estimer par $\widehat{v}_{\mathcal{M}}$ via la règle du maximum a posteriori :

$$\widehat{v}_{\mathcal{M}} \in \underset{v\in\mathcal{V}}{\operatorname{argmax}} \quad p(v|c, \widehat{\theta}_{\mathcal{M}}),$$

où $\widehat{\theta}_{\mathcal{M}}$ est une estimation de θ.

Enfin, le critère ICL sélectionne le modèle maximisant la log-vraisemblance complète intégrée en ayant remplacé v^* par $\widehat{v}$:

$$ICL(H, \mathcal{M}) = \log p(c, \widehat{v}_{\mathcal{M}}).$$

Par souci de concision, le critère $ICL(H, \mathcal{M})$ sera noté $ICL(H)$.

Remarque. Dans le chapitre suivant, nous verrons qu'au lieu de remplacer v^* par $\widehat{v}$, Wyse et al. (2016) proposent d'optimiser directement le critère en v afin d'estimer v^* et nous comparerons ces deux stratégies dans la section 3.2.4.

De manière générale, le calcul de ce critère peut être explicite pour certains choix de lois a priori, ce qui est le cas pour les lois a priori conjuguées que nous avons choisies.

Nous nous intéressons à l'adaptation du critère d'inspiration bayésienne ICL (*Integrated Completed Likelihood*) pour le LBM Poisson, étudié dans le cadre du modèle des blocs latents pour données catégorielles par Keribin et al. (2015). Dans cette étude, nous discutons également de l'influence des hyperparamètres sur le critère. Nous proposons enfin, à la manière de Keribin et al. (2015), une forme pour le critère BIC pour le modèle étudié.

4.1 Sélection de modèles pour le LBM Poisson normalisé

4.1.1 Critère Intregated Completed Likelihood (ICL)

Nous reprenons les lois a priori définies précédemment pour les paramètres du modèle des blocs latents pour les données de Poisson :

$$\rho \sim \mathcal{D}(a, .., a) \quad , \quad \tau \sim \mathcal{D}(a, .., a) \quad \text{et} \quad \forall h, \ell, \quad \gamma_{h\ell}|\alpha, \beta \sim \Gamma(\alpha, \beta).$$

Le critère $ICL(H, L)$ sélectionne alors le modèle maximisant la log-vraisemblance intégrée complète :

$$ICL(H, L) = \log p(c, \widehat{v}, \widehat{w}).$$

Pour calculer ce terme, nous utilisons l'hypothèse d'indépendance conditionnelle par rapport à θ de v et w :

$$
\begin{aligned}
& \int p(c,v,w|\gamma,\rho,\tau)\ p(\gamma)\ p(\rho)\ p(\tau)\ d\gamma\ \ d\rho\ d\tau \\
= & \int p(c|v,w,\gamma,\rho,\tau)\ p(v,w|\gamma,\rho,\tau)\ p(\gamma)\ p(\rho)\ p(\tau)\ d\gamma\ d\rho\ d\tau \\
= & \int p(c|v,w,\gamma,\rho,\tau)\ p(v|\gamma,\rho,\tau)\ p(w|\gamma,\rho,\tau)\ p(\gamma)\ p(\rho)\ p(\tau)\ d\gamma\ d\rho\ d\tau \\
= & \int p(c|v,w,\gamma)\ p(v|\rho)\ p(w|\tau)\ p(\gamma)\ p(\rho)\ p(\tau)\ d\gamma\ d\rho\ d\tau \\
= & \int p(c|v,w,\gamma)\ d\gamma \int p(v|\rho)\ p(\rho)\ d\rho \int p(w|\rho)\ p(\tau)\ d\tau \\
= & \ p(c|v,w)\ p(v)\ p(w).
\end{aligned}
$$

Nous obtenons ainsi la décomposition :

$$
ICL(H,L) = \log p(c|\widehat{v},\widehat{w}) + \log p(\widehat{v}) + \log p(\widehat{w}). \tag{4.1}
$$

Avec le choix précédent des lois a priori, nous obtenons une formule explicite (voir section 6.3.1 pour les calculs) du critère ICL pour le modèle Poisson LBM :

$$
\begin{aligned}
ICL(H,L) = & \log\Gamma(H\times a) + \log\Gamma(L\times a) - (H+L)\log\Gamma(a) + HL\left(\alpha\log\beta - \log\Gamma(\alpha)\right) \\
& - \log\Gamma(J + H\times a) - \log\Gamma(K + L\times a) - \sum_{j,k} \log c_{jk}! + c_{jk}\log\mu_j + c_{jk}\log\nu_k \\
& + \sum_h \log\Gamma(\widehat{v}_{\bullet h} + a) + \sum_\ell \log\Gamma(\widehat{w}_{\bullet \ell} + a) + \sum_{h,\ell} \log\Gamma\left(\alpha + \sum_{j,k} \widehat{v}_{jh}\widehat{w}_{k\ell}c_{jk}\right) \\
& + \sum_{h,\ell} - \left(\alpha + \sum_{j,k} \widehat{v}_{jh}\widehat{w}_{k\ell}c_{jk}\right) \log\left(\beta + \sum_{j,k} \widehat{v}_{jh}\widehat{w}_{k\ell}\mu_j\nu_k\right)
\end{aligned} \tag{4.2}
$$

où le couple $(\widehat{v},\widehat{w})$ dépend des nombres de composants H et L.

Remarque. Certains termes intervenant dans la formule du critère ne dépendent pas du nombre de composantes H et L et on peut ne pas les inclure en pratique.

Les estimateurs des partitions sont obtenus à partir de la règle du maximum a posteriori en utilisant l'estimation de θ fournie par l'algorithme *V-Bayes*.
L'influence des hyperparamètres sera discuté dans la section 4.2.3.a.

4.1.2 Critère Bayesian Information Criterion (BIC)

Le critère *Bayesian Information Criterion* (BIC) se place dans un contexte bayésien. Nous nous appuyons dans la suite sur la présentation faite par Lebarbier and Mary-Huard (2004). Pour ce faire, une collection finie de modèles est considérée, le paramètre θ_H et le modèle $\mathcal{M}_H$ sont des variables aléatoires et sont munies de lois a priori. L'avantage de cette méthode est qu'elle permet de prendre en compte des informations détenues par l'utilisateur et mettre un

poids plus important à certains modèles. Une première possibilité pour le choix du nombre de classes consiste à choisir le modèle $\widehat{\mathcal{M}}$ qui maximise la probabilité a posteriori :

$$\widehat{\mathcal{M}} \in \underset{\mathcal{M}_H}{\arg}\, p(\mathcal{M}_H|c).$$

En d'autres termes, BIC sélectionne ainsi le modèle le plus vraisemblable au vu des données c. En utilisant la formule de Bayes et le fait que la loi a priori sur les modèles $\mathcal{M}_h$ est non informative, la recherche du meilleur modèle ne nécessite que le calcul de la vraisemblance intégrée :

$$p(c;\mathcal{M}_H) = \int_{\theta_H} p(c;\mathcal{M}_H,\theta_h)p(\theta_H)d\theta_H.$$

où $p(\theta_H)$ est une distribution a priori non informative sur le paramètre θ_H. Mais le calcul de cette vraisemblance intégrée est souvent impossible. Le critère *Bayesian Information Criterion* (BIC), proposé par Schwarz (1978), est une approximation asymptotique du logarithme de la vraisemblance intégrée, effectuée à l'aide d'une approximation de Laplace (Lebarbier and Mary-Huard (2004)). Ce critère s'écrit :

$$BIC(H) = \log p(c;\mathcal{M}_H,\widehat{\theta}_H) - \frac{D_H}{2}\log J.$$

où θ_H est l'estimateur du maximum de vraisemblance des paramètres du modèle de mélange et D_H le degré de liberté du modèle à H composantes. Ce critère fournit une approximation pertinente de la vraisemblance lorsque le nombre d'observations J tend vers l'infini. Le critère sélectionne alors le modèle le plus proche de la loi des données au sens de la pseudo-distance de Kullback–Leibler $KL(f;g)$ entre deux lois de probabilité de densités respectives f et g Keribin (2000) :

$$KL(f,g) = \int \log\left(\frac{f}{g}\right) f(x)dx.$$

Mais, dans le cadre du modèle des blocs latents, la log-vraisemblance n'est pas calculable analytiquement, et donc aucun critère du type log-vraisemblance pénalisée n'est disponible.

Toutefois, une heuristique developpée par Keribin et al. (2015) permet malgré tout de proposer une expression du critère BIC.

En effet, elle repose d'une part, sur le développement asymptotique du critère ICL ci-dessous (Biernacki et al. (2000) pour les modèles de mélanges simples, Keribin et al. (2015) pour le LBM avec données catégorielles).

Une approximation asymptotique en J et en K d'ICL pour le LBM Poisson s'écrit :

$$ICL^{Asympt}(H,L) \approx \max_{\theta} \log p(c,\widehat{v},\widehat{w};\theta) - \frac{H-1}{2}\log J - \frac{L-1}{2}\log K - \frac{HL}{2}\log JK. \quad (4.3)$$

D'autre part, elle est basée sur une *conjecture* qui suppose que le terme d'entropie $\log p(\widehat{v},\widehat{w}|c;\widehat{\theta})$ intervenant dans la décomposition classique (Biernacki et al. (2000)) suivante, devient négligeable asymptotiquement lorsque nous nous plaçons sous le modèle qui a servi à générer les données :

$$ICL(H,L) = \log p(\widehat{v},\widehat{w}|c;\theta) + BIC(H,L). \quad (4.4)$$

Enfin, la dernière étape permettant d'aboutir à une expression pour le critère BIC est la suivante. Le paramètre maximisant la log-vraisemblance complète intégrée est remplacé par

l'estimateur du maximum de vraisemblance. Dans notre cas, comme cet estimateur n'est pas disponible, nous proposons de le remplacer par l'estimateur variationnel du maximum de l'énergie libre $\mathcal{F}$, soit l'estimateur $\widehat{\theta}^{VB}$ fourni par l'algorithme *V-Bayes*.

$$\begin{aligned}\max_{\theta} \log p(c, \widehat{v}, \widehat{w}; \theta) &\approx \log p(c, \widehat{v}, \widehat{w}; \widehat{\theta}) \\ &\approx \log p(\widehat{v}, \widehat{w} | c; \widehat{\theta}) + \log p(c; \widehat{\theta}) \\ &\approx \log p(\widehat{v}, \widehat{w} | c; \widehat{\theta}^{VB}) + \mathcal{F}(\widehat{\theta}^{VB})\end{aligned}$$

Ainsi, en utilisant 4.4 et 4.3, nous pouvons proposer une forme pour le critère BIC.
Pour le LBM Poisson, nous proposons une expression heuristique du critère $BIC(H, L)$:

$$BIC(H, L) = \mathcal{F}(\widehat{\theta}^{VB}) - \frac{HL + H - 1}{2} \log J - \frac{HL + L - 1}{2} \log K. \tag{4.5}$$

Nous pouvons observer de manière classique dans le modèle des blocs latents qu'il y a à la fois une pénalité sur le nombre de lignes et sur le nombre de colonnes, ce qui confirme la double asymptotique de ce modèle. Nous retrouvons alors la même forme du critère BIC proposé par Keribin et al. (2015)) dans le cas des données catégorielles.

4.2 Aspects algorithmiques : Procédure Bi-KM1

4.2.1 Initialisations récursives

Étant donné que nous devons parcourir un nombre de classes en ligne et aussi en colonne, le nombre de couples à explorer devient très grand comparé au cas d'un mélange classique unidimensionnel. Ainsi, il est encore possible d'effectuer une recherche exhaustive dans le cas de matrices de petites tailles, mais dans le cas de données massives, comme c'est le cas ici, celle-ci n'est pas réalisable en pratique.

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 *KM1* ("K minus 1") proposée dans le cas d'un mélange simple, par Baudry and Celeux (2015) et basée sur des initialisations récursives. En effet, si l'utilisateur vise à choisir un nombre de classes H variant dans l'ensemble $\{H_{min}, \ldots, H_{max}\}$, l'initialisation récursive consiste à découper au hasard une des H composantes en deux composantes afin d'obtenir la solution à $(H+1)$ composantes qui initialisera l'algorithme EM. Cette méthode d'initialisation, concurrente des méthodes classiques d'initialisation (initialisation aléatoire, small-EM, SEM, CEM, pour une revue détaillée voir Baudry and Celeux (2015)) a l'avantage d'éviter des estimations de paramètres non pertinents et fournit des graphiques avec une log-vraisemblance qui ne cesse d'augmenter. Nous utilisons alors cette idée qui nous fournit un cadre judicieux pour parcourir de manière non exhaustive le nombre de classes dans le cas bi-directionnel que nous impose le modèle des blocs latents.

4.2.2 Algorithme forward proposé : *Bi-KM1*

Supposons que nous devions choisir un modèle de mélange avec un nombre de composantes en ligne et en colonne appartenant respectivement à $\{H_{min}, \ldots, H_{max}\}$ et $\{L_{min}, \ldots, L_{max}\}$. L'initialisation récursive consiste à découper aléatoirement une des H ou des L composantes en deux afin d'obtenir le modèle $\mathcal{M}_{H+1,L}$ à $(H+1, L)$ composantes et le modèle $\mathcal{M}_{H,L+1}$ qui initialisera l'algorithme *V-Bayes*. Il convient de préciser que

- la solution (H_{min}, L_{min}) est obtenue minutieusement avec par exemple une procédure telle que l'échantillonneur de Gibbs couplé avec l'algorithme *V-Bayes* (Keribin et al. (2015)).
- Partant de $(H, L) = (H_{min}, L_{min})$, la position initiale du modèle $\mathcal{M}_{(H+1,L)}$ et celle du modèle $\mathcal{M}_{(H,L+\ 1)}$ est obtenue en découpant l'une des composantes du modèle $\mathcal{M}_{(H,L)}$ en deux et ce découpage est effectué de manière exhaustive sur chacune des composantes en ligne et en colonne (*Complete Choice*). Enfin, à partir de ces $H + L$ initialisations de l'algorithme, la solution retenue entre le modèle $\mathcal{M}_{(H+1,L)}$ et le modèle $\mathcal{M}_{(H,L+1)}$ 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 8.7), à la manière d'un algorithme *forward* classique. Remarquons qu'il n'est pas possible d'effectuer une étape *backward* dans cette procédure.

Ainsi, au lieu de parcourir une grille de taille $H_{max} \times L_{max}$, l'algorithme visite dans le pire des cas $H_{max} + L_{max}$ couples de classes en ligne et en colonne.

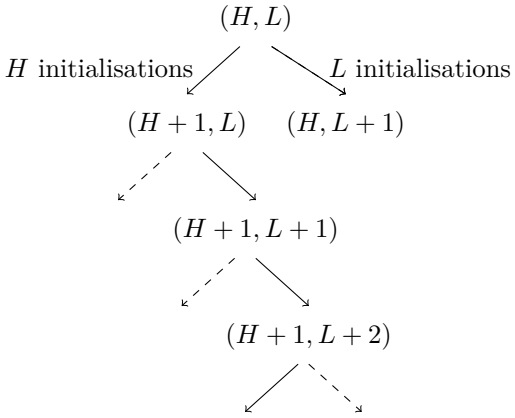


Figure 4.1 – Représentation schématique de l'algorithme *Bi-KM1*.

4.2.3 Expérimentations numériques

4.2.3.a Discussion autour des hyperparamètres

Pour tester l'influence des hyperparamètres sur le critère ICL, nous reprenons les trois scénarios générant différents types de matrices de la section 3.2.3.b pour une taille de matrices égale à $(J, K) = (500, 500)$ et un nombre de classes en ligne et en colonne $(H, L) = (4, 5)$.

Nous utilisons la procédure *Bi-KM1* pour des classes en lignes et en colonne allant de 2 à 8, puis nous regardons le couple des nombres sélectionné par ICL suivant les valeurs des hyperparamètres en faisant varier ces derniers dans l'ensemble $\{10^{-10}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Pour chaque configuration, sont simulées 100 matrices.

Les résultats sont résumés dans les tableaux 4.1, 4.2 et 4.3 qui représentent respectivement les scénarios 1, 2 et 3. Nous observons que le critère ICL semble peu sensible aux hyperparamètres, ce qui est très satisfaisant. De plus, il sélectionne dans la plupart des cas le bon nombre de classe en lignes et en colonne. À noter que lorsque $\delta > 1$, le critère ICL sélectionne parfois un nombre de classes en ligne et en colonne inférieur à celui qui a servi à

générer les données. Au vu des résultats, et par continuité du chapitre 2, nous proposons de prendre les mêmes hyperparamètres utilisés pour l'estimation, à savoir $\delta = 1$ et $\zeta = 0.01$.

	$\zeta = 10^{-10}$			$\zeta = 0.001$			$\zeta = 0.01$			$\zeta = 0.1$			$\zeta = 1$			$\zeta = 10$		
$\delta = 10^{-10}$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.001$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.01$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 10$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100

Table 4.1 – Sur 100 matrices, répartition du nombre de classes obtenu par la procédure *Bi-KM1* suivant la valeur de l'hyperparamètre δ (lignes globales) et la valeur de l'hyperparamètre ζ (colonnes globales), pour $\varepsilon = 0$ (scénario 1).

	$\zeta = 10^{-10}$			$\zeta = 0.001$			$\zeta = 0.01$			$\zeta = 0.1$			$\zeta = 1$			$\zeta = 10$		
$\delta = 10^{-10}$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.001$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.01$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 10$		4	5		4	5		4	5		4	5		4	5		4	5
	3	2	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	98	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100

Table 4.2 – Sur 100 matrices, répartition du nombre de classes obtenu par la procédure *Bi-KM1* suivant la valeur de l'hyperparamètre δ (lignes globales) et la valeur de l'hyperparamètre ζ (colonnes globales), pour $\varepsilon = 2$ (scénario 2) .

	$\zeta = 10^{-10}$			$\zeta = 0.001$			$\zeta = 0.01$			$\zeta = 0.1$			$\zeta = 1$			$\zeta = 10$		
$\delta = 10^{-10}$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.001$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.01$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 0.1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 1$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100	4	0	100
$\delta = 10$		4	5		4	5		4	5		4	5		4	5		4	5
	3	0	0	3	23	0	3	5	0	3	0	0	3	0	0	3	0	0
	4	0	100	4	0	77	4	0	95	4	0	100	4	0	100	4	0	100

Table 4.3 – Sur 100 matrices, répartition du nombre de classes obtenu par la procédure *Bi-KM1* suivant la valeur de l'hyperparamètre δ (lignes globales) et la valeur de l'hyperparamètre ζ (colonnes globales), pour $\varepsilon = 3$ (scénario 3).

4.2.3.b Intérêt de la normalisation : paramètres μ et ν

Il s'agit ici de comparer les deux modélisations possibles pour le modèle des blocs latents dans le cas de données de comptage, à savoir $\mathcal{P}(\lambda)$ (sans normalisation : modélisation 1) contre $\mathcal{P}(\mu\nu\gamma)$ (avec normalisation : modélisation 2) en utilisant la procédure *Bi-KM1*.

Pour ce faire, nous avons simulé dans un premier temps, cinquante matrices issues de la modélisation sans normalisation (modélisation 1) avec les paramètres suivants : $(J, K) = (1000, 1000)$, $(H, L) = (4, 5)$, $\rho_h = \frac{h}{10}, h = 1, \ldots, 4$, $\tau_\ell = \frac{\ell}{15}, \ell = 1, \ldots, 5$ et :

$$\lambda = \begin{pmatrix} 7 & 6 & 5 & 5 & 5 \\ 6 & 7 & 6 & 5 & 5 \\ 5 & 6 & 7 & 6 & 5 \\ 5 & 5 & 6 & 7 & 6 \end{pmatrix}.$$ (correspondant au cas ε=4 de la section 4.2.5 b.)

Un exemple d'une telle matrice est présentée dans la figure 4.2. Nous avons également représenté les histogrammes des marges lignes et marges colonnes de la matrice. Comme elle a été simulée selon la modélisation sans normalisation, nous pouvons remarquer que la distribution des marges lignes et colonnes au sein de chaque classe se concentre autour de sa moyenne (pour plus de précisions, voir *Algorithme Largest Gaps*, Channarond et al. (2012)).

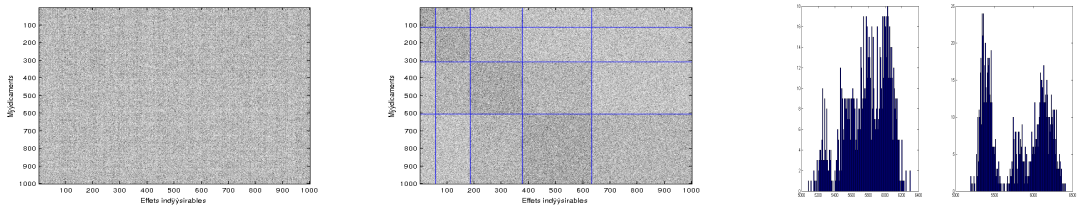
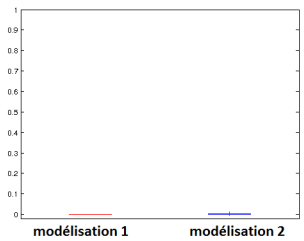


Figure 4.2 – Exemple de matrice simulée (à gauche), réorganisée en partitions (au milieu) ainsi que les histogrammes de ses marges lignes et colonnes (à droite)

Nous avons appliqué la procédure *KM1* en utilisant chacune des deux modélisations et la figure 4.3 montre que l'erreur de classification croisée est très satisfaisante pour les deux types de modélisation, avec une très légère supériorité de la procédure avec la modélisation 1, car nous y perdons un peu à estimer les paramètres μ et ν. De plus, pour les deux types de modélisation le critère ICL sélectionne à coup sûr le bon nombre de classes.



ICL, modélisation 1

h \ ℓ	4	5	6
4	0	50	0
5	0	0	0
6	0	0	0

ICL, modélisation 2

h \ ℓ	4	5	6
4	0	50	0
5	0	0	0
6	0	0	0

Figure 4.3 – Erreur de classification croisée sur 50 matrices (à gauche) et répartition du nombre de classes en ligne et en colonne sélectionné par la procédure *Bi-KM1* (à droite) selon le type de modélisation.

Dans un second temps, nous avons simulé cinquante matrices issues de la modélisation avec normalisation (modélisation 2) avec les paramètres suivants : $(J, K) = (1000, 1000)$, $(H, L) = (4, 5)$, $\rho_h = \frac{h}{10}, h = 1, \ldots, 4$, $\tau_\ell = \frac{\ell}{15}, \ell = 1, \ldots, 5$ et :

$$\lambda = \begin{pmatrix} 7 & 6 & 5 & 5 & 5 \\ 6 & 7 & 6 & 5 & 5 \\ 5 & 6 & 7 & 6 & 5 \\ 5 & 5 & 6 & 7 & 6 \end{pmatrix}$$ (correspondant au cas ε=4 de la section 4.2.5 b.).

Un exemple d'une telle matrice est présentée dans la figure 4.4. Nous avons également représenté les histogrammes des marges lignes et marges colonnes de la matrice. Contrairement à la matrice précédente, nous n'observons plus ce phénomène de concentration.

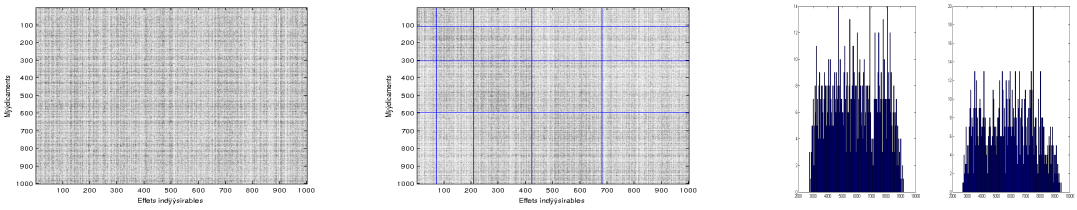
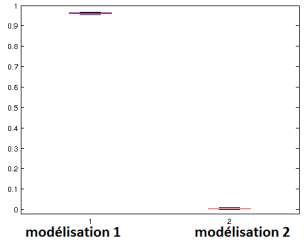


Figure 4.4 – Exemple de matrice simulée (à gauche), réorganisée en partitions (au milieu) ainsi que les histogrammes de ses marges lignes et colonnes (à droite)



ICL, modélisation 1

h \ ℓ	4	5	6
4	0	0	0
5	0	0	0
6	0	0	0

ICL, modélisation 2

h \ ℓ	4	5	6
4	0	50	0
5	0	0	0
6	0	0	0

Figure 4.5 – Erreur de classification croisée sur 50 matrices (à gauche) et répartition du nombre de classes en ligne et en colonne sélectionné par la procédure *Bi-KM1* (à droite) selon le type de modélisation.

Nous avons appliqué de même la procédure *KM1* en utilisant chacune des deux modélisations et la figure 4.5 montre que la procédure implémentée avec la modélisation 1 échoue à retrouver la classification croisée de matrices issues de la modélisation 2, et à tendance surestimer le nombre de classes en ligne et colonne.

Ainsi, la procédure *KM1* implémentée avec la modélisation avec normalisation (modélisation 2) réalise de bonnes performances à la fois sur des matrices issues de la modélisation 1 ou 2.

4.2.3.c Variabilité de l'algorithme

Comme la procédure *Bi-KM1* repose sur des tirages aléatoires lors des découpages de classes en deux, nous allons tester empiriquement sa variabilité (couples de classes en ligne et en colonne visités, couple de classes sélectionnés). Pour ce faire, nous éprouvons la procédure

sur une simulation de matrice plus grande, avec un nombre de classes en ligne et en colonne très différent l'un par rapport à l'autre et plus important que ce qui était présenté dans les plans d'expériences précédents.

La matrice considérée (voir figure 4.6) est de taille $(J, K) = (1000, 1000)$ simulée avec un nombre de classes en ligne et en colonne $(H, L) = (10, 20)$. Les paramètres μ et ν sont simulés suivant une loi uniforme $\mathcal{U}_{(0.5,1.5)}$ et le paramètre γ suivant une loi uniforme $\mathcal{U}_{(0,10)}$.

L'expérience consiste à réaliser dix exécutions indépendantes de la procédure sur la matrice considérée. Dès que l'une des composantes d'un couple visité dépasse la taille de la grille fixée $(12, 22)$, la procédure s'arrête. Enfin, sur les couples visités, le critère ICL et le critère BIC sont calculés et le couple sélectionné correspond à la valeur la plus élevée du critère.

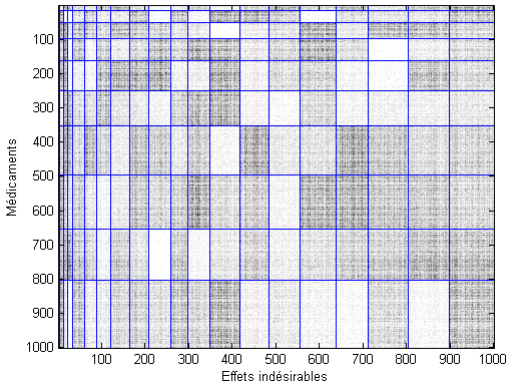


Figure 4.6 – Représentation de la matrice de données utilisée pour mesurer la variabilité de la procédure *Bi-KM1*. Cette matrice est réorganisée par les partitions ayant servi à la simulation avec $(J, K) = (1000, 1000)$ et $(H, L) = (10, 20)$.

Dans le tableau 4.4 sont représentés les couples sélectionnés par le critère ICL et le critère BIC. Pour chacune des dix exécutions, la procédure *Bi-KM1* a non seulement visité le bon nombre de classes en ligne et en colonne mais l'a également sélectionné pour les deux critères.

H \ L	19	20	21
9	0	0	0
10	0	10	0
11	0	0	0

H \ L	19	20	21
9	0	0	0
10	0	10	0
11	0	0	0

Table 4.4 – Répartition du nombre de classes en ligne H et en colonne L obtenu par ICL (à gauche) et par BIC (à droite) durant 10 exécutions.

Nous avons également présenté dans la figure 4.7 les trajectoires des couples visités par chacune des dix exécutions de la procédure *Bi-KM1* et nous observons que les couples visités au cours de la procédure varient beaucoup mais que toutes les exécutions visitent le couple de composantes qui a servi à générer les données. Enfin, nous pouvons observer que la procédure a visité en moyenne 29 couples, ce qui est nettement moins que les $12 \times 22 = 264$ couples présents dans la grille.

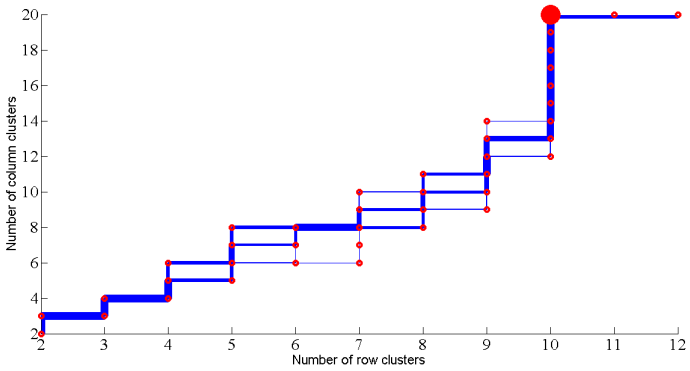


Figure 4.7 – Trajectoires des couples visités par chacune des dix exécutions.

4.2.4 Comparaison avec un algorithme de recherche gloutonne pour optimiser ICL

Les deux procédures que nous allons comparer n'ont à l'origine pas le même objectif. En effet, la procédure *Bi-KM1* vise à estimer les paramètres d'un modèle en utilisant un estimateur du mode a posteriori, puis maximise le critère ICL par rapport au nombre de classes en ligne et en colonne. La procédure que nous allons détailler dans ce qui suit et à laquelle nous allons nous comparer est la procédure de recherche gloutonne (GS, Wyse et al. (2016)). Contrairement à la première procédure, elle vise à optimiser ICL en les partitions et intègre donc le critère sur l'espace des paramètres. Nous allons voir que la procédure *Bi-KM1*, même si ce n'est pas son objectif premier permet d'obtenir des partitions autant satisfaisantes que celles fournis par la procédure GS.

Afin d'effectuer une comparaison des procédures, nous nous plaçons dans le modèle des blocs latents de Poisson avec la modélisation sans normalisation $\mathcal{P}(\lambda)$, modèle étudié par Wyse et al. (2016).

4.2.4.a Algorithme de recherche gloutonne (GS, Wyse et al. (2016))

Dans le cadre du modèle des blocs latents de Poisson, nous rappelons que le critère ICL s'écrit :

$$ICL(H, L) \quad = \quad \log p(v) + \log p(w) + \log p(c|v, w)$$

avec

$$p(v) = \frac{\Gamma(aH)}{\Gamma(a)^H} \frac{\prod_h \Gamma(v_{\cdot h} + a)}{\Gamma(J + aH)},$$

$$p(w) = \frac{\Gamma(aL)}{\Gamma(a)^L} \frac{\prod_\ell \Gamma(w_{\cdot \ell} + a)}{\Gamma(K + aL)},$$

et

$$p(c|v, w) = \frac{p(c|v, w, \lambda)p(\lambda)}{p(\lambda|c, v, w)}.$$

L'algorithme de recherche gloutonne pour ICL (GS, Wyse et al. (2016)) utilise une approche similaire à Malsiner-Walli et al. (2016) pour choisir les nombres de classes H et L.

Premièrement, ils initialisent les labels v et w en choisissant des valeurs conservatives (plus grandes que nécessaire) que nous notons H_{max} et L_{max}.

L'algorithme GS affecte itérativement les observations aux classes et fusionnent les classes existantes afin de maximiser le critère ICL.

Plus précisément, l'algorithme balaie de manière aléatoire les lignes. Il choisit alors une ligne telle que $v_{jh} = 1$. Puis il calcule la différence $\Delta_{h\to h'}$ observée dans le calcul d'ICL en déplaçant la ligne j dans une classe h' pour tout $h' \neq h$:

$$\Delta_{h\to h'} = ICL(H, L, \tilde{v}, w) - ICL(H, L, v, w),$$

où $\tilde{v}_{rh'} = v_{rh'}$, $\forall r \neq j$ et $\tilde{v}_{jh'} = 1$.

Enfin, la ligne j est affectée à la classe h' qui donne la plus grande valeur positive de $\Delta_{h\to h'}$. Si $\Delta_{h\to h'} < 0$, $\forall h'$, la ligne j reste dans sa classe originelle.

Si le déplacement de la ligne j de la classe h entraîne que cette classe soit vide, la différence calculée est alors

$$\Delta_{h\to h'} = ICL(H - 1, L, \tilde{v}, w) - ICL(H, L, v, w).$$

C'est ce processus qui permet de vider les classes au fur et à mesure que la recherche gloutonne progresse.

Ce processus est également effectué de manière indépendante sur les colonnes et la recherche s'arrête lorsque qu'aucun déplacement de lignes ou de colonnes n'entraîne une augmentation de $\Delta_{h\to h'}$ pour les lignes ou $\Delta_{\ell\to \ell'}$ pour les colonnes.

4.2.4.b Expérimentations numériques sur données simulées

Les données ont été générées suivant le modèle des blocs latents avec $H = 4$ et $K = 5$. Les proportions de mélange en ligne et en colonne ont été choisies déséquilibrées, suivant une progression arithmétique, c'est-à-dire $\rho_h = \frac{h}{10}, h = 1, \ldots, 4$ et $\tau_\ell = \frac{\ell}{15}, \ell = 1, \ldots, 5$. Le nombre de lignes et de colonnes varient parmi les couples suivants $(J, K) = (50, 50), (100, 100), (500, 500)$. Utilisant la modélisation de Poisson sans normalisation ($\mathcal{P}(\lambda)$), le paramètre λ dépend d'un paramètre ε que nous faisons varier entre 3 et 4 pour obtenir trois scénarios (voir figure 4.8) matérialisant différents niveaux de chevauchement entre les blocs :

$$\lambda = \begin{pmatrix} 15-2\varepsilon & 10-\varepsilon & 5 & 5 & 5 \\ 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon & 5 & 5 \\ 5 & 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon & 5 \\ 5 & 5 & 10-\varepsilon & 15-2\varepsilon & 10-\varepsilon \end{pmatrix}.$$

Ainsi, plus ε se rapproche de la valeur 5, plus des ex-aequo sont générés et les classes en ligne et en colonne sont de moins en moins séparées, la tâche de classification devient alors plus difficile. Cent jeux de données ont été simulés pour chaque valeur de ε correspondant à un scénario.

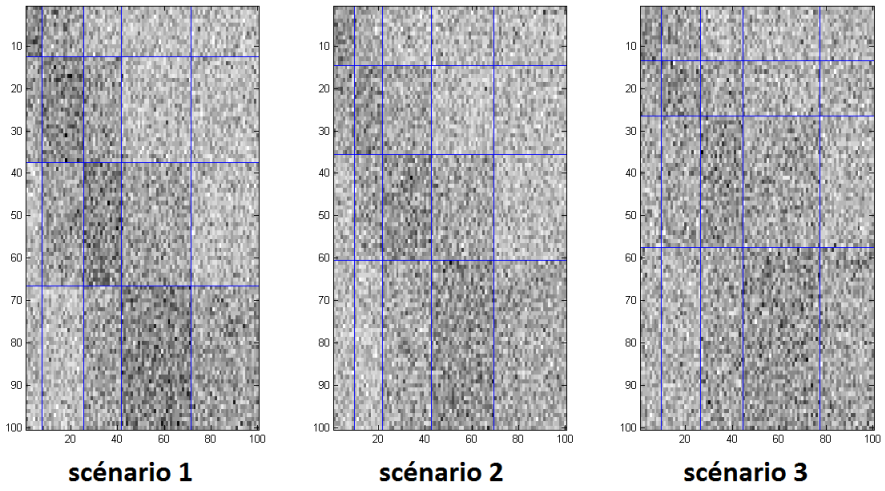


Figure 4.8 – Exemple de matrices simulées c de données de comptage de tailles $(100, 100)$ et réorganisées pour les trois scénarios considérés pour $\varepsilon = 3$, 3.5 et 4, de gauche à droite.

Concernant l'initialisation de la procédure *GS*, aucune suggestion n'est donnée dans Wyse et al. (2016) et nous choisissons l'option par défaut.

Les deux procédures préconisent les mêmes valeurs pour les hyperparamètres δ et ζ, à savoir $\delta = 1$ et $\zeta = 0.01$ mais ont des préférences assez distinctes concernant l'hyperparamètre a. Pour la procédure *Bi-KM1*, il est préconisé d'utiliser $a = 4$ pour éviter le phénomène de dégénérescence des classes alors que pour la procédure *GS*, il est préféré un a plus petit (les travaux de Malsiner-Walli et al. (2016) préconisent $a = 0.1$ pour permettre que les classes superflues se vident rapidement), mais remarquons que dans le cas de la procédure *GS*, la valeur par défaut choisie est $a = 1$.

Nous allons tester les deux procédures sur chacun de leur terrain de prédilection.

La procédure *GS* [*] est optimisée et en général plus rapide pour l'instant que la procédure *Bi-Km1* qui sera optimisée par la suite. De plus, nous voyons que la procédure *Bi-Km1* est assez stable vis-à-vis de l'initialisation où le modèle des blocs latents a servi à générer les données et quelle que soit la taille des données. En effet, pour chaque jeu d'hyperparamètres, nous avons effectué 10 exécutions des deux procédures et représenté la boxplot des valeurs maximales d'ICL pour chacune des 100 matrices du scénario 3. Pour la procédure *GS*, il existe une variabilité quant à la valeur maximale d'ICL renvoyée, surtout lorsque la taille des données augmente (voir figures 4.9, 4.10, 4.11, 4.12, 4.13 et 4.14). Partant de ces constats, nous choisissons de comparer une exécution de la procédure *Bi-KM1* contre dix lancements de la procédure *GS*.

Plus précisément, les approches considérées sont les suivantes :

- une exécution de la procédure *Bi-KM1* initialisée avec une solution présentant $(H, L) = (2, 2)$ classes en ligne et colonne obtenue par le couplage des algorithmes *Gibbs+V-Bayes*,
- dix exécutions de la procédure *GS* initialisées depuis une solution aléatoire présentant $(H, L) = (20, 20)$ classes en ligne et en colonne.

[*] https://sites.google.com/site/jsnwyse/code

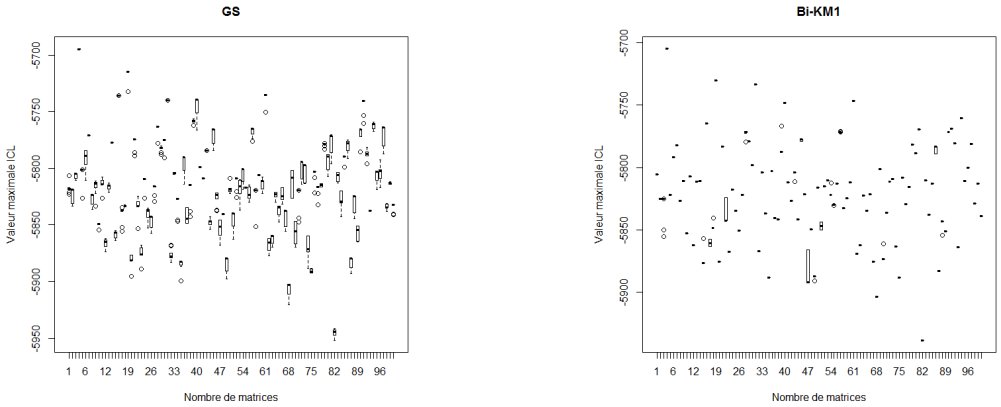


Figure 4.9 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(50, 50)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(4, 1, 0.01)$.

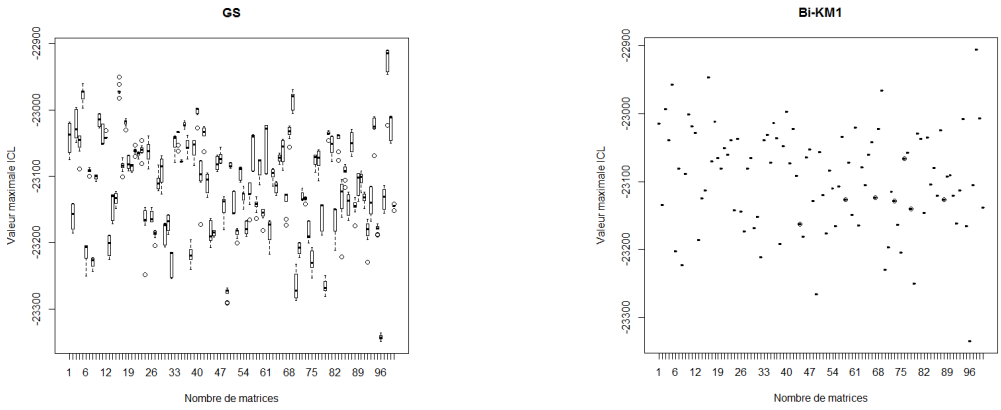


Figure 4.10 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(100, 100)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(4, 1, 0.01)$.

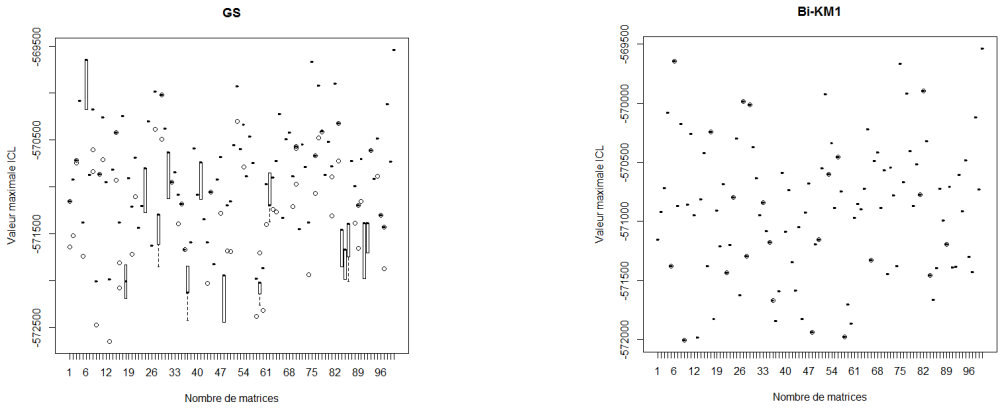


Figure 4.11 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(500, 500)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(4, 1, 0.01)$.

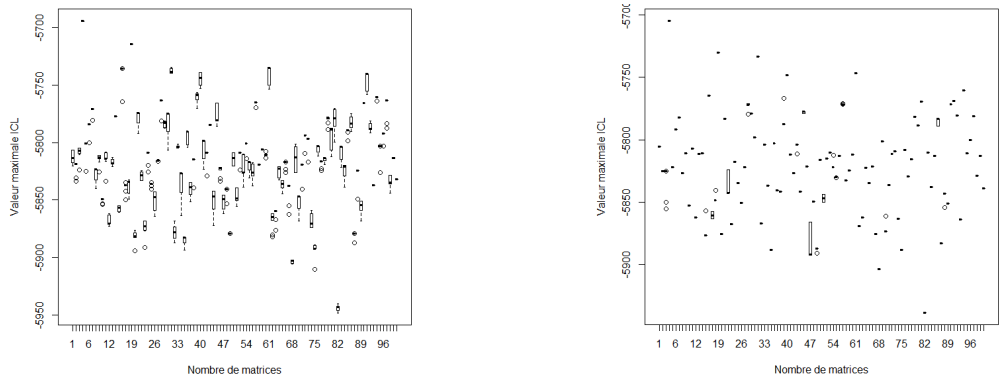


Figure 4.12 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(50, 50)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(1, 1, 0.01)$.

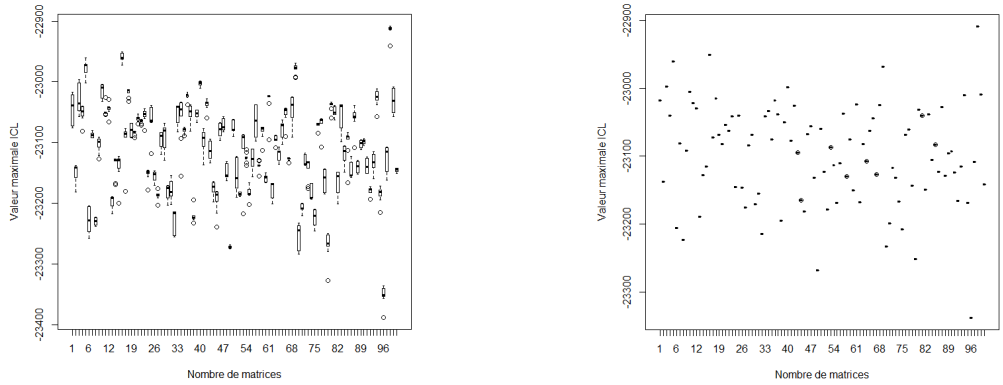


Figure 4.13 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(100, 100)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(1, 1, 0.01)$.

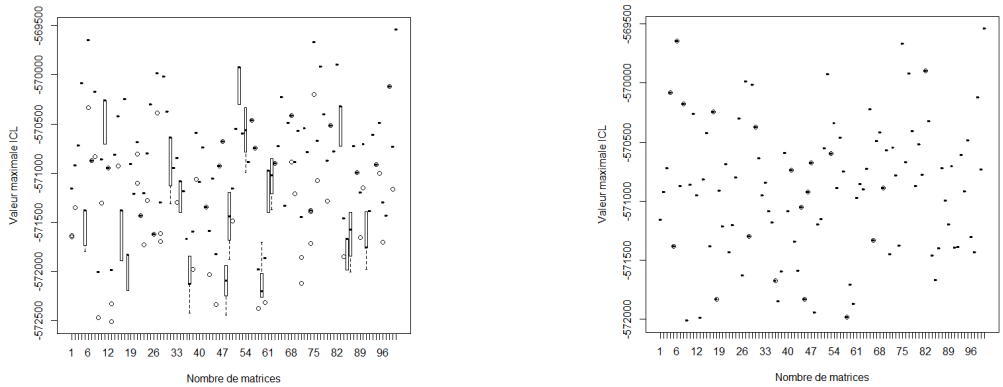


Figure 4.14 – Boxplot des valeurs maximales d'ICL renvoyées par dix exécutions de la procédure *GS* (à gauche) et dix exécutions de la procédure *Bi-KM1* (à droite) pour chacune des 100 matrices provenant du scénario 3 de taille d'échantillon $(500, 500)$ pour le jeu d'hyperparamètres (a, δ, ζ)=$(1, 1, 0.01)$.

Comparaison pour les hyperparamètres $(4, 1, 0.01)$

Ici, les approches que nous comparons ont été testées pour le même jeu d'hyperparamètres.

Pour chaque couple final de partitions obtenues par chacune des procédures, le critère ICL a été calculé.

Les figures 4.15 et 4.16 présentent les résultats montrant la comparaison des valeurs maximales d'ICL calculées pour les deux procédures, pour chacun des 3 scénarios lorsque la taille d'échantillon est petite $(J, K) = (50, 50)$ et grande $(J, K) = (500, 500)$. Le cas modéré où $(J, K) = (100, 100)$ sera détaillé plus loin.

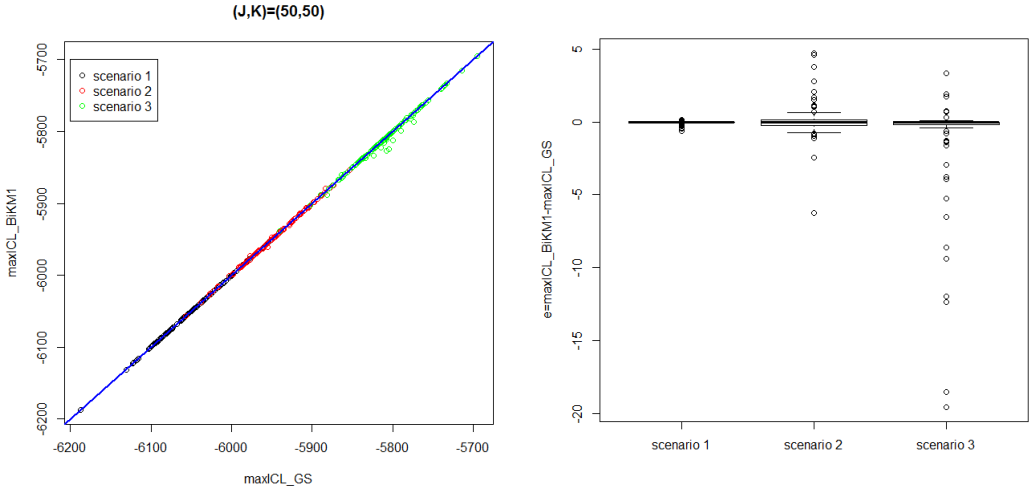


Figure 4.15 – Comparaison des valeurs maximales d'ICL renvoyées par la procédure *GS* (en abscisse) et par la procédure *Bi-KM1* (en ordonnée) et boxplot des différences des valeurs renvoyées par les procédures suivant 3 scénarios pour la taille d'échantillon : $(50, 50)$.

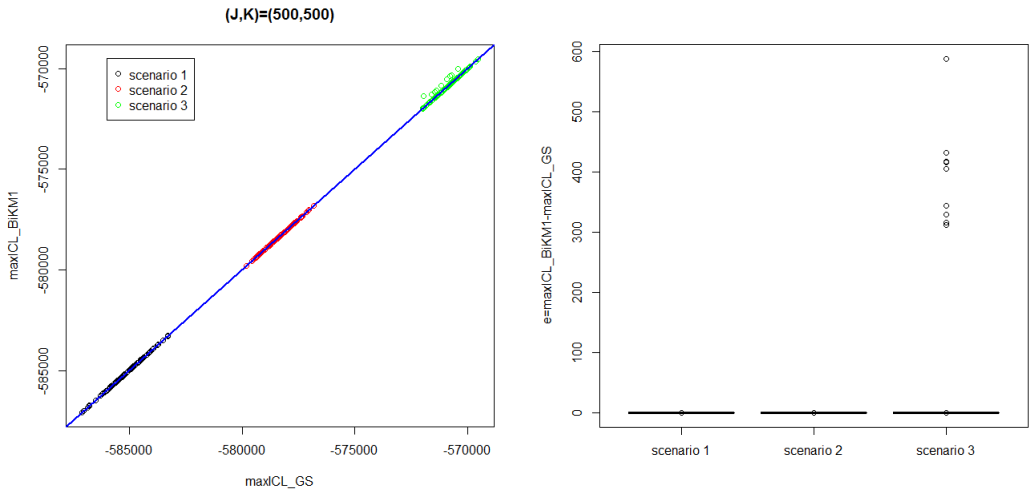


Figure 4.16 – Comparaison des valeurs maximales d'ICL renvoyées par la procédure *GS* (en abscisse) et par la procédure *Bi-KM1* (en ordonnée) et boxplot des différences des valeurs renvoyées par les procédures suivant 3 scénarios pour la taille d'échantillon : $(500, 500)$.

Pour de petites tailles d'échantillon $(J, K) = (50, 50)$, la procédure *GS* produit de meilleures performances pour le scénario 3 (difficile).

Dans le cas où nous avons des grandes tailles d'échantillon $(J, K) = (500, 500)$, les performances de la procédure *GS* et la procédure *Bi-KM1* sont similaires pour le scénario 1 et 2. Quant au scénario 3 (difficile), les performances de la procédure *Bi-KM1* sont plus satisfaisantes. Ainsi, pour certaines matrices, dix exécutions de la procédure *GS* ne sont peut-être pas suffisantes pour obtenir la valeur maximale d'ICL.

Étude détaillée du cas modéré : $(J, K) = (100, 100)$

La figure 4.17 présente les résultats montrant la comparaison des valeurs maximales d'ICL renvoyées par les deux procédures, pour chacun des 3 scénarios dans le cas où la taille des données est modérée $(J, K) = (100, 100)$. Dans ce cas-ci, les performances des deux algorithmes sont similaires

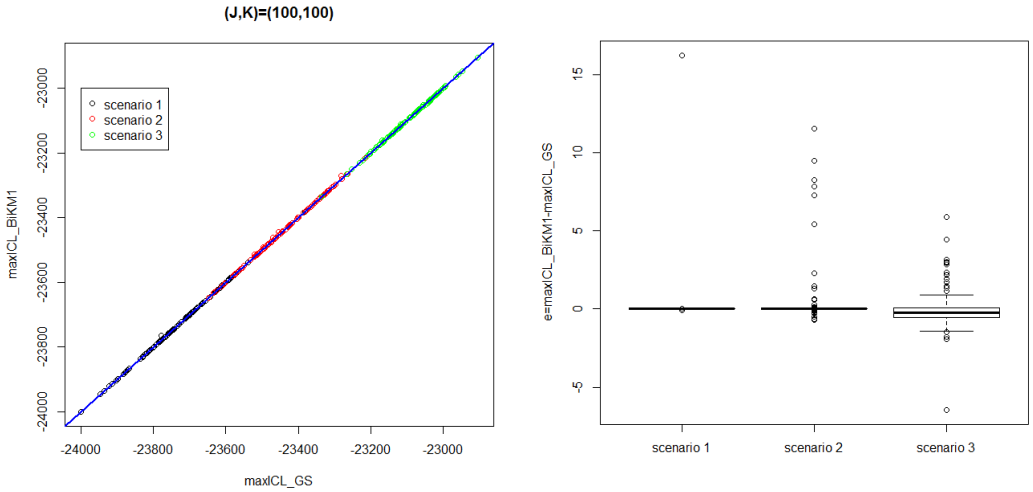
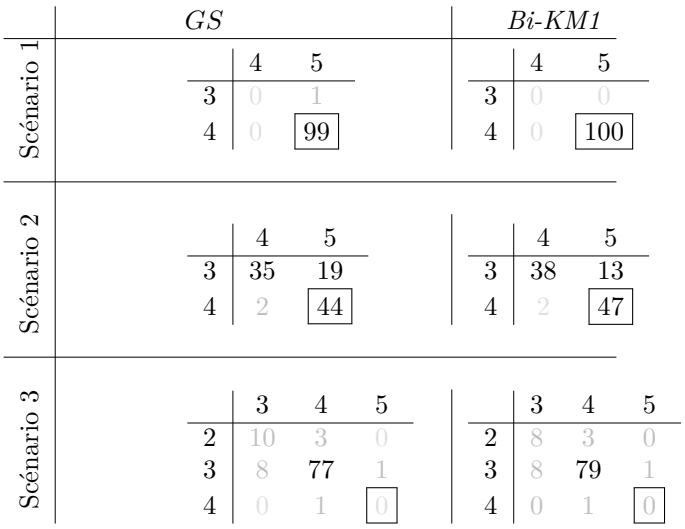


Figure 4.17 – Comparaison des valeurs maximales d'ICL renvoyées par la procédure *GS* (en abscisse) et par la procédure *Bi-KM1* (en ordonnée) et boxplot des différences des valeurs renvoyées par les procédures suivant 3 scénarios pour la taille d'échantillon : $(100, 100)$.

Le tableau 4.5 présente la répartition du nombre de classes en ligne et en colonne renvoyé par les deux procédures (colonnes globales) sur 100 matrices simulées pour chacun des 3 scénarios (lignes globales). La procédure *Bi-KM1* réalise légèrement de meilleures performances et lorsque la difficulté des matrices augmentent, les deux procédures ont tendance à sous-estimer le nombre de classes en lignes et en colonne, ce qui est logique.

Afin de comparer plusieurs classifications d'un jeu de données, nous considérons l'erreur de classification introduite par Lomet (2012) et le *Coclustering Adjusted Rand Index* présentées dans la section 3.3.3, qui permet d'évaluer deux classifications croisées entre elles. L'erreur de Lomet (2012) vaut 0 si les partitions sont les mêmes à une permutation près et le *Coclustering Adjusted Rand Index* vaut 1 si les partitions sont les mêmes à une permutation près. Les performances des deux procédures présentées dans le tableau 4.6 et 4.7 sont comparables, avec une légère supériorité de la procédure *BI-KM1* pour les deux erreurs.



Scénario 1 — GS	4	5
3	0	1
4	0	**99**

Scénario 1 — Bi-KM1	4	5
3	0	0
4	0	**100**

Scénario 2 — GS	4	5
3	35	19
4	2	**44**

Scénario 2 — Bi-KM1	4	5
3	38	13
4	2	**47**

Scénario 3 — GS	3	4	5
2	10	3	0
3	8	77	1
4	0	1	**0**

Scénario 3 — Bi-KM1	3	4	5
2	8	3	0
3	8	79	1
4	0	1	**0**

Table 4.5 – Répartition du nombre de classes en ligne et en colonne renvoyé par les deux procédures (colonnes globales) sur 100 matrices simulées pour chacun des 3 scénarios (lignes globales). Le rectangle représente le nombre de classes en ligne et en colonne qui a servi à générer les données.

	GS	*Bi-KM1*
Scénario 1	0.0063 (0.0134)	0.0052 (0.0069)
Scénario 2	0.1011 (0.0779)	0.0970 (0.0785)
Scénario 3	0.3264 (0.1129)	0.3108 (0.1087)

Table 4.6 – Moyenne sur 100 matrices simulées de l'erreur de classification de Lomet (2012) pour les deux procédures en comparaison avec les labels qui ont servi à générer les données, suivant les trois scénarios. Entre parenthèses, est énoncé l'écart-type.

	GS	*Bi-KM1*
Scénario 1	0.9924 (0.0122)	0.9929 (0.0109)
Scénario 2	0.9105 (0.0554)	0.9125 (0.0581)
Scénario 3	0.6150 (0.1273)	0.6364 (0.1231)

Table 4.7 – Moyenne sur 100 matrices simulées du *Coclustering Adjusted Rand Index* pour les deux procédures en comparaison avec les labels qui ont servi à générer les données, suivant les trois scénarios. Entre parenthèses, est énoncé l'écart-type.

Comparaison pour les hyperparamètres $(1, 1, 0.01)$

Ici, les approches que nous comparons ont été testées pour le même jeu d'hyperparamètres, intervenant dans le critère ICL : (a, δ, ζ)=$(1, 1, 0.01)$. La procédure *Bi-KM1* est plus sensible à la valeur de l'hyperparamètre a et dans le cas $(J, K) = (50, 50)$ pour le scénario 3, la procédure a éprouvé quelques difficultés car a été confronté au problème de dégénérescence des classes.

Les résultats étant similaires dans le cas où $(J, K) = (50, 50)$ et $(J, K) = (500, 500)$ nous présentons uniquement le cas détaillé où $(J, K) = (100, 100)$.

Étude détaillée du cas modéré : $(J, K) = (100, 100)$

La figure 4.18 présente les résultats montrant la comparaison des valeurs maximales d'ICL renvoyées par les deux procédures, pour chacun des 3 scénarios dans le cas où la taille des données est modérée $(J, K) = (100, 100)$. Les performances des deux algorithmes sont également comparables.

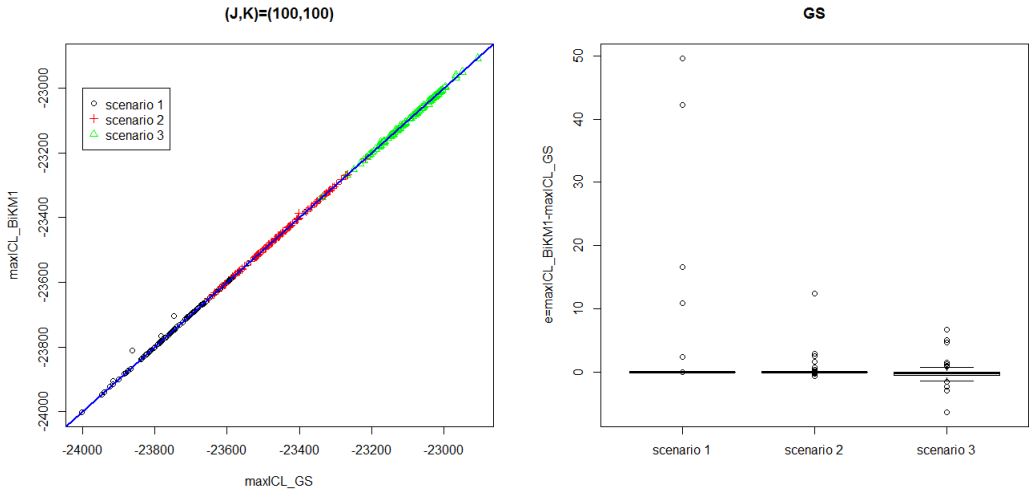


Figure 4.18 – Comparaison des valeurs maximales d'ICL renvoyées par la procédure *GS* (en abscisse) et par la procédure *Bi-KM1* (en ordonnée) et boxplot des différences des valeurs renvoyées par les procédures suivant 3 scénarios pour la taille d'échantillon : $(100, 100)$.

Le tableau 4.8 présente la répartition du nombre de classes en ligne et en colonne renvoyé par les deux procédures (colonnes globales) sur 100 matrices simulées pour chacun des 3 scénarios (lignes globales). La procédure *Bi-KM1*, même si les hyperparamètres utilisés ne sont pas ceux préconisés, réalise légèrement de meilleures performances et lorsque la difficulté des matrices augmentent, les deux procédures ont tendance à sous-estimer le nombre de classes en lignes et en colonne.

	GS			*Bi-KM1*		
Scénario 1		4	5		4	5
	3	1	3	3	0	0
	4	1	[95]	4	0	[100]
Scénario 2		4	5		4	5
	0	35	15	0	31	14
	4	2	[48]	4	2	[53]
Scénario 3		3	4	5	3	4	5
	2	12	2	0	2	9	3	0
	3	8	77	1	3	7	79	1
	4	0	0	[0]	4	0	1	[0]

Table 4.8 – Répartition du nombre de classes en ligne et en colonne renvoyé par les deux procédures (colonnes globales) sur 100 matrices simulées pour chacun des 3 scénarios (lignes globales). Le rectangle représente le nombre de classes en ligne et en colonne qui a servi à générer les données.

Les performances des deux procédures en terme de classification croisée sont présentées dans les tableaux 4.9 et 4.10 et les erreurs les plus faibles pour chacun des 3 scénarios sont réalisées pour la procédure *Bi-KM1*.

	GS	*Bi-KM1*
Scénario 1	0.0103 (0.0256)	0.0049 (0.0066)
Scénario 2	0.0968 (0.0794)	0.0882 (0.0768)
Scénario 3	0.3294 (0.1170)	0.3129 (0.1094)

Table 4.9 – Moyenne sur 100 matrices simulées de l'erreur de classification de Lomet (2012) pour les deux procédures en comparaison avec les labels qui ont servi à générer les données, suivant les trois scénarios. Entre parenthèses, est énoncé l'écart-type.

	GS	*Bi-KM1*
Scénario 1	0.9902 (0.0169)	0.9932 (0.0066)
Scénario 2	0.9129 (0.0573)	0.9184 (0.0555)
Scénario 3	0.6132 (0.1312)	0.6335 (0.1216)

Table 4.10 – Moyenne sur 100 matrices simulées du *Coclustering Adjusted Rand Index* pour les deux procédures en comparaison avec les labels qui ont servi à générer les données, suivant les trois scénarios. Entre parenthèses, est énoncé l'écart-type.

Conclusion

La procédure *GS* semble être sensible aux initialisations, ce qui s'accentue lorsque les tailles d'échantillons sont grandes, mais demeure pour l'instant plus rapide. De plus, aucune préconisation ou précision n'est fournie sur la stratégie d'initialisation et sur le choix des hyperparamètres, qui peut influencer le phénomène qui permet de vider les classes superflues.

La procédure *Bi-KM1*, quant à elle est moins sensible aux initialisations dans le cas où le "vrai" modèle appartient à la collection considérée, mais elle est plus rigide concernant le jeu d'hyperparamètres utilisé. Après étude de sensibilité, le jeu d'hyperparamètres préconisé est $(4, 1, 0.01)$. Par ailleurs, elle peut être parallélisée pour gagner en rapidité et semble être légèrement plus performante pour trouver le bon nombre de classes lorsque le modèle des blocs latents a servi à générer les données.

Ainsi, même si l'objectif de la procédure *Bi-KM1* n'est pas d'optimiser le critère ICL en les partitions, elle réalise de bonnes performances en classification croisée.

4.2.4.c Application aux données réelles "MovieLens dataset"

Modélisation sans normalisation

Les jeux de données intitulés "MovieLens" sont librement accessibles depuis Internet [†] et plusieurs tailles de données y sont également disponibles. Dans cette étude, nous nous concentrons sur le jeu de données comprenant 100 000 classements de films allant de 1 à 5 effectués par 943 utilisateurs impliquant 1682 films. La matrice des données considérée est représentée sur la figure 4.20 et notons que lorsqu'une donnée est manquante, elle est représentée par la valeur 0.

L'hypothèse faite par Wyse et al. (2016) est que les données sont distribuées selon une loi de Poisson, hypothèse justifiée par leur caractère ordinal et discret. De plus, les auteurs précisent que cette modélisation permettrait de glaner plus d'information que si les données avaient été traitées de manière catégorielle. En revanche, une perspective intéressante serait de comparer les résultats de cette modélisation avec celle de Biernacki and Jacques (2015) qui elle prend bien en compte le caractère ordinal des données.

Ainsi, l'objectif de classification dans lequel nous nous plaçons est de proposer des groupes d'individus présentant une manière similaire de noter des groupes spécifiques de films, ce qui permet par exemple de proposer des systèmes de recommandations.

Pour réaliser cette tâche, nous allons utiliser les deux procédures présentées précédemment en utilisant les mêmes hyperparamètres, à savoir $(a = 1, \delta = 1, \zeta = 1)$.

La solution retenue pour la procédure *GS* nous a été fournie par les auteurs. La solution retenue pour la procédure *Bi-KM1* correspond à une exécution de celle-ci pour le critère ICL et BIC.

Les résultats sont présentés dans le tableau 4.13. Les valeurs maximales d'ICL sont globalement du même ordre de grandeur, de même que le couple de classes en ligne et en colonne pour tous les critères.

L'erreur de Lomet (2012) n'étant clairement pas calculable en pratique étant donné que le nombre de classes en ligne ou en colonne est largement supérieur à 8, nous avons calculé l'*Adjusted Rand Index étendu* entre les différentes partitions fournis par la procédure *GS* d'une part et la procédure *Bi-KM1* avec le critère ICL et BIC d'autre part. Les partitions fournies par le critère BIC et le critère ICL sont celles qui présentent le meilleur indice (voir tableau 4.12).

[†] http://grouplens.org/datasets/movielens/

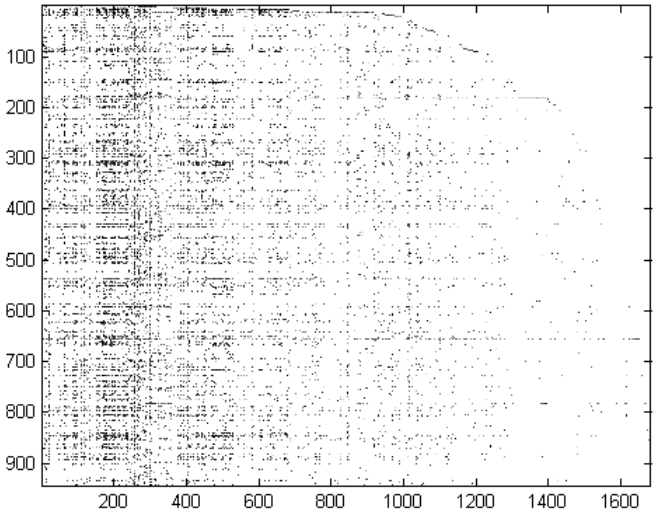


Figure 4.19 – Données MovieLens comportant 100000 classements d'utilisateurs allant de 1 à 5 impliquant 943 utilisateurs et 1682 films.

	GS, *ICL*	*Bi-KM1*, *ICL*	*Bi-KM1*, *BIC*
Valeur du Critère	-649 358.1	-648 368.8	-380 094.29
$\widehat{H}$ (classe en ligne)	50	56	54
$\widehat{L}$ (classe en colonne)	55	49	52

Table 4.11 – Comparaison des deux procédures sur les données réelles MovieLens.

	GS, *ICL*	*Bi-KM1*, *ICL*	*Bi-KM1*, *BIC*
GS, *ICL*	1	0.3513	0.3559
Bi-KM1, *ICL*	0.3513	1	0.4895
Bi-KM1, *BIC*	0.3559	0.4895	1

Table 4.12 – Comparaison des partitions proposées par les procédures via le *Coclustering Adjusted Rand Index* sur les données réelles MovieLens.

Modélisation avec normalisation

Nous voulons tester notre procédure en utilisant la modélisation avec normalisation ($\mathcal{P}(\mu\nu\gamma)$) et voir si elle fournit un nombre de classes en ligne et en colonne du même ordre de grandeur que ceux présentés précédemment avec la modélisation sans normalisation.

Pour justifier le choix de la modélisation, une indication intéressante est de regarder l'histogramme des marges lignes et colonnes de la matrice des données. Nous n'observons aucun phénomène de concentration, donc la modélisation avec normalisation semble plus indiquée. De plus, adopter la modélisation avec normalisation signifierait que deux utilisateurs qui auraient la même manière de noter les mêmes films mais avec un décalage proportionnel des notes peuvent être mis dans la même classe, ce qui n'est pas le cas avec la modélisation sans normalisation.

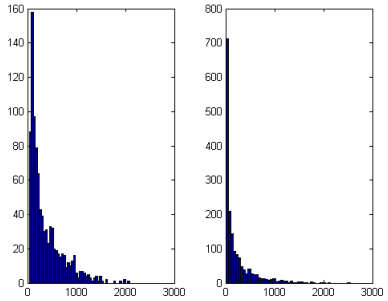


Figure 4.20 – Histogramme des marges lignes (à gauche) et des marges colonnes (à droite) de la matrice de données.

Les résultats d'une exécution de la procédure *Bi-KM1* avec la modélisation avec normalisation sont présentés dans les tableaux 4.13 et 4.14 pour différents jeux d'hyperparamètres. Nous voyons que le choix d'hyperparamètres influence peu les résultats obtenus de la procédure. Nous remarquons, comme attendu, que le nombre de classes en ligne et en colonne obtenu que ce soit par le critère ICL est nettement moindre que les résultats précédents avec la modélisation sans normalisation, ce qui peut être plus appréciable pour analyser les classes obtenues.

	Bi-KM1, ICL
Valeur du Critère	-4 743 694.5
$\widehat{H}$ (classe en ligne)	32
$\widehat{L}$ (classe en colonne)	21

Table 4.13 – Résultat d'une exécution de la procédure *Bi-KM1* avec normalisation pour le jeu d'hyperparamètres $(4, 1, 0.01)$.

	Bi-KM1, ICL
Valeur du Critère	-4 743 527.8
$\widehat{H}$ (classe en ligne)	31
$\widehat{L}$ (classe en colonne)	21

Table 4.14 – Résultat d'une exécution de la procédure *Bi-KM1* avec normalisation pour le jeu d'hyperparamètres $(1, 1, 0.01)$.

4.3 Annexes

4.3.1 Détails de la preuve de la formule 4.2

Nous devons calculer chacun des trois termes apparaissant dans 4.1:

- Commençons par le calcul de $p(v)$. On sait que conditionnellement à ρ, la loi de v est une loi multinomiale de paramètres 1 et ρ. D'où :

$$p(v|\rho) = \prod_{j,h} \rho_h^{v_{jh}}.$$

De plus comme la loi a priori de ρ est une loi de Dirichlet $\mathcal{D}(a, .., a)$:

$$p(\rho) = \frac{\Gamma(Ha) \prod_h \rho_h^{a-1}}{\Gamma(a)^H},$$

nous obtenons :

$$p(\rho|v) = \frac{p(v|\rho)p(\rho)}{\int p(v|\rho)p(\rho)d\rho} \propto \prod_h \rho_h^{v_{\bullet h}+a-1},$$

avec $v_{\bullet h} = \sum_{j=1}^{J} v_{jh}$. Nous reconnaissons alors une densité non normalisée de Dirichlet $\mathcal{D}(v_{\bullet 1} + a, .., v_{\bullet H} + a)$ dont la normalisation vaut :

$$\frac{\Gamma(J + Ha)}{\prod_h \Gamma(v_{\bullet h} + a)}.$$

Ainsi d'après la formule de Bayes, nous avons finalement :

$$p(v) = \frac{\Gamma(aH)}{\Gamma(a)^H} \frac{\prod_h \Gamma(v_{\bullet h} + a)}{\Gamma(J + aH)}.$$

D'où en passant au logarithme, on obtient

$$\log p(v) = \log \Gamma(aH) + \sum_{h=1}^{H} [\log \Gamma(v_{\bullet h} + a)] - H \log \Gamma(a) - \log \Gamma(J + aH).$$

- De la même manière, nous avons :

$$p(w) = \frac{\Gamma(aL)}{\Gamma(a)^L} \frac{\prod_\ell \Gamma(w_{\bullet \ell} + a)}{\Gamma(K + aL)},$$

et ainsi,

$$\log p(w) = \log \Gamma(aL) + \sum_{\ell=1}^{L} [\log \Gamma(w_{\bullet \ell} + a)] - L \log \Gamma(a) - \log \Gamma(K + aL),$$

- Il ne reste plus qu'à expliciter ce terme :

$$
\begin{aligned}
\log p(c|v,w) &= \log p(c|v,w,\gamma) + \log p(\gamma) - \log p(\gamma|c,v,w) \\
&= \sum_{j,k,h,\ell} v_{jh} w_{k\ell} \log \varphi\left(c_{jk}; \mu_j \nu_k \gamma_{h\ell}\right) + \sum_{h,\ell} \left(\log p(\gamma_{h\ell}) - \log p(\gamma_{h\ell}|c,v,w)\right) \\
&= \sum_{h,\ell} \left[\alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log \gamma_{h\ell} - \beta \gamma_{h\ell}\right] \\
&\quad + \sum_{j,k,h,\ell} v_{jh} w_{k\ell} \left[-\mu_j \nu_k \gamma_{h\ell} + c_{jk} \log \gamma_{h\ell} + c_{jk} \log \mu_j + c_{jk} \log \nu_k - \log c_{jk}!\right] \\
&\quad - \sum_{h,\ell} \left[\left(\alpha + \sum_{jk} v_{jh} w_{k\ell} c_{jk}\right) \log \left(\beta + \sum_{j,k} v_{jh} w_{k\ell} \mu_j \nu_k\right) - \log \Gamma \left(\alpha + \sum_{jk} v_{jh} w_{k\ell} c_{jk}\right) \right. \\
&\quad \left. + \left(\alpha - 1 + \sum_{j,k} v_{jh} w_{k\ell} c_{jk}\right) \log \gamma_{h\ell} - \gamma_{h\ell} \left(\sum_{j,k} v_{jh} w_{k\ell} \mu_j \nu_k + \beta\right) \right]
\end{aligned}
$$

Et finalement, nous obtenons en utilisant le fait que $\sum_h v_{jh} = \sum_\ell w_{k\ell} = 1$,

$$
\begin{aligned}
\log p(c|v,w) &= HL\left(\alpha \log \beta - \log \Gamma(\alpha)\right) + \sum_{j,k} c_{jk} \log \mu_j + c_{jk} \log \nu_k - \log c_{jk}! \\
&\quad + \sum_{h,\ell} \left[- \left(\alpha + \sum_{j,k} v_{jh} w_{k\ell} c_{jk}\right) \log \left(\beta + \sum_{j,k} v_{jh} w_{k\ell} \mu_j \nu_k\right) + \log \Gamma \left(\alpha + \sum_{j,k} v_{jh} w_{k\ell} c_{jk}\right) \right]
\end{aligned}
$$

4.3.2 Détails de la preuve de la formule 4.3

Une approximation asymptotique de chaque terme peut être effectuée dans l'égalité ci-dessous :

$$
ICL(H,L) = \log p(c, \widehat{v}, \widehat{w}) = \log p(c|\widehat{v}, \widehat{w}) + \log p(\widehat{v}) + \log p(\widehat{w}) \tag{4.6}
$$

Biernacki et al. (2000) proposent d'utiliser une approximation de la fonction Gamma :

$$
\Gamma(z) \underset{z \to +\infty}{\sim} z^{z - \frac{1}{2}} e^{-z} \sqrt{2\pi},
$$

soit

$$
\log \Gamma(z) \underset{z \to +\infty}{\sim} (z - \frac{1}{2}) \log z - z - \frac{1}{2} \log 2\pi.
$$

Afin de pouvoir utiliser l'approximation de la fonction Gamma pour chacun des termes dans 4.6, il faut supposer que $\widehat{v_{\cdot h}} \underset{J \to +\infty}{\to} +\infty$ et $\widehat{w_{\cdot \ell}} \underset{K \to +\infty}{\to} +\infty$.

On a alors

$$
\log p(\widehat{v}) = \log \Gamma(aH) + \underbrace{\sum_{h=1}^{H} [\log \Gamma(\widehat{v}_{\cdot h} + a)]}_{A} - H \log \Gamma(a) - \underbrace{\log \Gamma(J + aH)}_{B}.
$$

Si on néglige alors tous les termes ne dépendant pas de J, nous avons,

$$\begin{aligned}
A &\sim \sum_h [(\widehat{v_{\bullet h}} + a - \frac{1}{2}) \log(\widehat{v_{\bullet h}} + a) - (\widehat{v_{\bullet h}} + a)] \\
&\sim \sum_h \widehat{v_{\bullet h}} \log \left(\widehat{v_{\bullet h}} (1 + \frac{a}{\widehat{v_{\bullet h}}}) \right) + (a - \frac{1}{2}) \log \widehat{v_{\bullet h}} - \widehat{v_{\bullet h}} - a \\
&\sim \sum_h \widehat{v_{\bullet h}} \log \widehat{v_{\bullet h}} + \cancel{\widehat{v_{\bullet h}} \frac{a}{\widehat{v_{\bullet h}}}} + (a - \frac{1}{2}) \log \widehat{v_{\bullet h}} - \widehat{v_{\bullet h}} - \cancel{a} \\
&\sim \sum_h \widehat{v_{\bullet h}} \log \widehat{v_{\bullet h}} - J + (a - \frac{1}{2}) \log \widehat{v_{\bullet h}} \\
&\sim \sum_h [\widehat{v_{\bullet h}} \log \widehat{v_{\bullet h}}] - J + H(a - \frac{1}{2}) \log J,
\end{aligned}$$

car $\widehat{\rho_h} = \frac{\widehat{v_{\bullet h}}}{J}$.

De plus,

$$\begin{aligned}
B &\sim (J + Ha - \frac{1}{2}) \log(J + Ha) - (J + Ha) \\
B &\sim J[\log J + \frac{Ha}{J}] + (Ha - \frac{1}{2})[\log J + \frac{Ha}{J}] - J - Ha \\
B &\sim J \log J + (Ha - \frac{1}{2}) \log J - J - Ha,
\end{aligned}$$

car on peut négliger $\frac{Ha}{J}$.

D'où

$$A - B \sim \sum_h \widehat{v_{\bullet h}} \log \widehat{v_{\bullet h}} - \cancel{J} + \cancel{H(a} - \frac{1}{2}) \log J - J \log J - \cancel{(Ha} - \frac{1}{2}) \log J + \cancel{J} + Ha$$

Or $\sum_h \widehat{v_{\bullet h}} \log \widehat{v_{\bullet h}} - J \log J = \max_\rho \log p(\widehat{v}; \rho)$ donc finalement on conclut que

$$\log p(\widehat{v}) \sim \max_\rho \log p(\widehat{v}; \rho) - \frac{H - 1}{2} \log J,$$

et de même,

$$\log p(\widehat{w}) \sim \max_\tau \log p(\widehat{w}; \tau) - \frac{L - 1}{2} \log K.$$

Maintenant regardons,

$$\begin{aligned}\log p(c|\widehat{v},\widehat{w}) &= HL\left(\alpha\log\beta-\log\Gamma(\alpha)\right)+\sum_{j,k}c_{jk}\log\mu_j+c_{jk}\log\nu_k-\log c_{jk}!\\ &\quad+\sum_{h,\ell}\left[-\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\log\left(\beta+\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k\right)+\log\Gamma\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\right]\end{aligned}$$

Ainsi, si nous négligeons tous les termes ne dépendant pas de J et K, nous avons,

$$\begin{aligned}&\log p\left(c\,|v,w\,\right)\\ &\underset{+\infty}{\sim}\underbrace{\sum_{j,k}\log\frac{\left(\mu_j\nu_k\right)^{c_{jk}}}{c_{jk}!}}_{A}+\underbrace{\sum_{h,\ell}\left[-\alpha\log\left(\beta+\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k\right)-\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\log\left(\beta+\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k\right)\right.}_{B}\\ &+\underbrace{\sum_{h,\ell}\log\Gamma\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)}_{C}\end{aligned}$$

Puis nous avons

$$A \sim \sum_{j,k}\sum_{h,\ell}v_{jh}w_{k\ell}\log\frac{\left(\mu_j\nu_k\right)^{c_{jk}}}{c_{jk}!}$$

et

$$B \sim -\sum_{h,\ell}\left[\alpha\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)-\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\log\left(\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k\right)\right]$$

$$
\begin{aligned}
C &\sim \sum_{h,\ell}\left[\log\left(\sqrt{2\pi}\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)^{\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}-1/2} e^{-\alpha-\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}\right)\right]\\
C &\sim \sum_{h,\ell}\left[\frac{1}{2}\log(2\pi)+\log\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)^{-1/2+\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}} - \log e^{\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}\right]\\
C &\sim \sum_{h,\ell}\left[-\frac{1}{2}\log\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)+\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\log\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)-\left(\alpha+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\log e\right]\\
C &\sim \sum_{h,\ell}\left[-\frac{1}{2}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)+\alpha\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)-\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right]\\
C &\sim \sum_{h,\ell}\left[-\frac{1}{2}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)+\alpha\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\right]\\
&\quad - \sum_{h,\ell}\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\frac{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}
\end{aligned}
$$

En rassemblant les termes A, B et C nous obtenons :

$$
\begin{aligned}
A+B+C &\sim \sum_{j,k}\sum_{h,\ell}v_{jh}w_{k\ell}\log\frac{(\mu_j\nu_k)^{c_{jk}}}{c_{jk}!}+\sum_{h,\ell}\left[\alpha\log\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}\right)+\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\log\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}\right)\right.\\
&\qquad \left.-\frac{1}{2}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\right]-\sum_{h,\ell}\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\frac{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}\\
&\sim \sum_{h,\ell}\left[\alpha\log\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}\right)-\frac{1}{2}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\right]\\
&\qquad \underbrace{\sum_{h,\ell}\sum_{j,k}v_{jh}w_{k\ell}\left[-\mu_j\nu_k\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}\right)+c_{jk}\log\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}\right)+\log\frac{(\mu_j\nu_k)^{c_{jk}}}{c_{jk}!}\right]}_{=\max_{\gamma}\log p(c|v,w;\gamma)}\\
&\underset{+\infty}{\sim} \max_{\gamma}\log p\left(c\,|v,w;\gamma\right)\\
&\qquad +\sum_{h,\ell}\left[\alpha\log\left(\frac{\sum_{j,k}v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k}v_{jh}w_{k\ell}\mu_j\nu_k}\right)-\frac{1}{2}\log\left(\sum_{j,k}v_{jh}w_{k\ell}c_{jk}\right)\right]
\end{aligned}
$$

Or, nous avons asymptotiquement :

$$\pi_{\min}\rho_{\min}\gamma_{\min}JK \leq \sum_{j,k}v_{jh}w_{k\ell}c_{jk} \leq \pi_{\max}\rho_{\max}\gamma_{\max}JK \text{ presque sûrement}$$

ce qui implique que :

$$\log\left(\pi_{\min}\rho_{\min}\gamma_{\min}\right)+\log(JK) \leq \log\left(\sum_{j,k} v_{jh}w_{k\ell}c_{jk}\right) \leq \log\left(\pi_{\max}\rho_{\max}\gamma_{\max}\right)+\log\left(JK\right) \text{ presque sûrement}$$

et donne donc que

$$\log\left(\sum_{j,k} v_{jh}w_{k\ell}c_{jk}\right) \underset{+\infty}{\sim} \log\left(JK\right) \text{ presque sûrement.}$$

Au final, nous avons :

$$\begin{aligned}
&\log p\left(c\,|v,w\right)\\
&\underset{+\infty}{\sim} \max_{\gamma} \log p\left(c\,|v,w;\gamma\right) + \sum_{h,\ell}\left[\alpha\log\left(\frac{\sum_{j,k} v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k} v_{jh}w_{k\ell}\mu_j\nu_k}\right) - \frac{1}{2}\log\left(JK\right)\right]\\
&\underset{+\infty}{\sim} \max_{\gamma} \log p\left(c\,|v,w;\gamma\right) - \frac{HL}{2}\log\left(JK\right) + \underbrace{\sum_{h,\ell}\left[\alpha\log\left(\frac{\sum_{j,k} v_{jh}w_{k\ell}c_{jk}}{\sum_{j,k} v_{jh}w_{k\ell}\mu_j\nu_k}\right)\right]}_{\text{négligeable car terme borné}}\\
&\underset{+\infty}{\sim} \max_{\gamma} \log p\left(c\,|v,w;\gamma\right) - \frac{HL}{2}\log\left(JK\right).
\end{aligned}$$

Nous concluons en utilisant le fait que :

$$\begin{aligned}
\max_{\theta} \log p(c,\widehat{v},\widehat{w};\theta) &= \max_{\rho} \log p(\widehat{v};\rho) + \max_{\tau} \log p(\widehat{w};\tau)\\
&+ \max_{\gamma} \log p(c|\widehat{v},\widehat{w};\gamma)
\end{aligned}$$

5

Application au tableau de contingence de pharmacovigilance

Nous remercions l'équipe $B2PHI$ de l'UMR 1181 (Inserm, Villejuif) pour leur collaboration et avoir rendu disponible les données de pharmacovigilance que nous allons traiter.

5.1 Statistiques descriptives préliminaires

5.1.1 Description du tableau de contingence

Les données collectées entre 2000 et 2010 impliquent **219 340** individus, **2142** classes de médicaments et **4216** classes d'événements indésirables. À partir des matrices de données individuelles x et y, nous rappelons que nous pouvons construire le tableau de contingence c (voir figure 5.1) via la relation suivante

$$c = t(x) \times y,$$

où t désigne l'opération transposée.

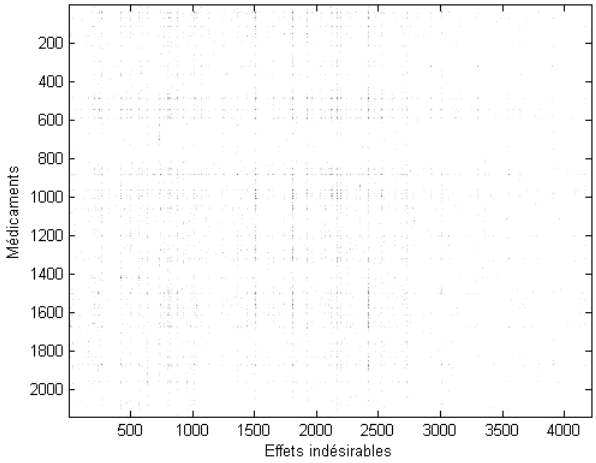


Figure 5.1 – Représentation du tableau de contingence de pharmacovigilance.

Le tableau suivant résume toutes les informations relatives au tableau de contingence c.

	table de contingence
Taille	$2142 \times 4216 = 9\ 030\ 672$
Valeur maximale présente	1684
Nombre de cellules non nulles	238 330 (2.64%)

Nous pouvons remarquer que les données sont massives et très parcimonieuses. L'histogramme des marges lignes et colonnes de la matrice des données est présenté dans la figure 5.2 et nous observons une grande dispersion des valeurs pour chacune des marges, ce qui justifie par la suite l'utilisation du modèle LBM normalisé.

De plus, un élément important à noter est que dans cette base, il n'existe pas de couples de médicaments et effets bi-univoques, c'est-à-dire un effet qui serait notifié qu'avec ce médicament et un médicament qui serait notifié qu'avec cet effet. Nous notons là qu'il existe une dissymétrie dans la matrice que nous souhaitons analyser.

5.1.2 Analyse Factorielle des Correspondances (AFC)

Le test du Chi-deux sur le tableau de contingence pour tester l'indépendance des médicaments et effets indésirables rejette, comme attendu en présence d'un échantillon de grande taille, l'hypothèse d'indépendance. Nous allons ensuite effectuer une analyse factorielle des correspondances. Les 2 premiers axes représentent environ 4,5% de l'inertie totale, ce qui est assez courant lorsque ce sont des données massives (voir figures 5.3 et 5.4).

L'AFC permet d'identifier des couples particuliers et nous remarquons que deux couples portent l'axe 1 et 2 et correspondent à des couples impliquant un effet rare et un médicament rarement pris dans la base, et notifiés ensemble. Ceux-ci correspondent à des couples *presque bi-univoques* très faiblement notifiés (ici généralement notifié une ou deux fois ensemble et notifié rarement avec d'autres variables). Remarquons que si :

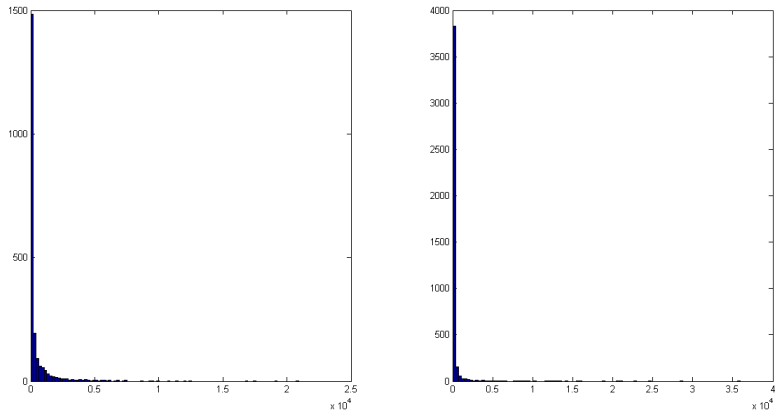


Figure 5.2 – Histogramme des marges lignes (à gauche) et des marges colonnes (à droite) de la matrice de données.

• Si le médicament existe depuis longtemps, cela peut vouloir dire qu'il a peu d'effets secondaires ou qu'il est peu pris. Dans les deux cas, ce couple ne sera peut-être pas une priorité.

• En revanche, si le médicament est récent, ces notifications sont peut-être les prémisses d'une relation avérée entre ce médicament et l'effet. Une surveillance de ce couple dans le futur est préconisée.

En revanche, aucun des couples mentionnés dans l'ensemble de référence OMOP décrit dans la section 1.1.3 n'est mis en évidence. Les premiers signaux mis en évidence par les méthodes usuelles décrites dans le chapitre 2, ne sont également pas les signaux référencés de l'OMOP. Ce fait peut s'expliquer par le fait que ce jeu de référence n'est pas exhaustif et d'autres signaux plus marqués mais non référencés sont mis en exergue par ces méthodes.

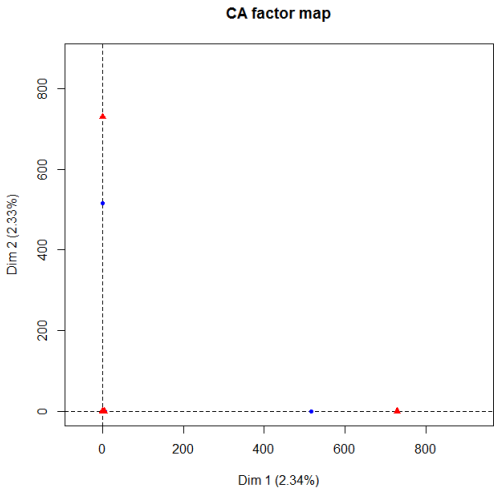


Figure 5.3 – Représentation simultanée des deux nuages profils lignes (médicaments symbolisés par des points) et profils colonnes (effets indésirables symbolisés par des triangles) dans le premier plan de l'AFC.

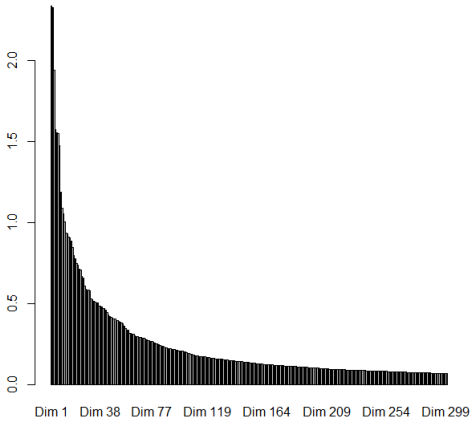


Figure 5.4 – Eboulis des valeurs propres en %.

En enlevant ces deux couples pour effectuer l'AFC et en réitérant ce processus jusqu'à qu'il n'y ait plus d'autres outliers que nous avons répertoriés dans le tableau 5.1, nous obtenons la 5.5 sur laquelle nous pouvons observer un effet taille. L'AFC ne semble pas alors le meilleur choix d'analyse.

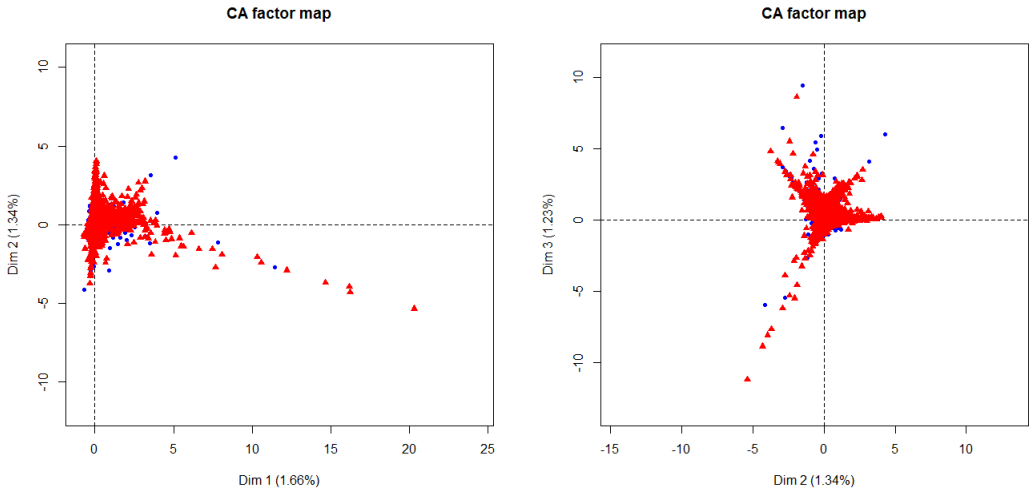


Figure 5.5 – Représentation simultanée des deux nuages profils lignes (médicaments symbolisés par des points) et profils colonnes (effets indésirables symbolisés par des triangles) dans le premier plan factoriel (à gauche) et deuxième plan factoriel (à droite) de l'AFC.

	Médicament (code ATC)	Effet (code Meddra)
Couple 1	L03AB09	10012758
Couple 2	A03AD01	10015120
Couple 3	B05ZB	10005706
Couple 4	A07AXXX	10003094
Couple 5	A14AA05	10057644
Couple 6	S01	10045210

Table 5.1 – Liste des couples presque bi-univoques, valeurs extrêmes de l'AFC.

5.2 Classification croisée du tableau de contingence

5.2.1 Élaboration d'une solution initiale pour la procédure *Bi-KM1*

5.2.1.a Étude préliminaire d'un tableau de contingence réduit

Dans un premier temps, nous pouvons nous restreindre à des notifications spontanées particulières, qui se révèlent une première source d'information intéressante, qui n'est souvent pas mis en exergue par les différentes méthodes existantes en pharmacovigilance. Nous allons construire un tableau de contingence réduit en ne prenant en compte que les individus qui ont pris un seul médicament et ont eu un seul effet, l'effet de coprescription est alors directement écarté de cette étude : $\tilde{c} = \tilde{x} \times \tilde{y}$.

Dans la base actuelle, cela représente 51 367 individus soit 20% des individus au total et concerne 1482 des 2142 médicaments et 2239 des 4216 effets indésirables (voir figure 5.6). Ici, la valeur $c_{\tilde{j}k}$ du tableau de contingence réduit correspond bien au nombre d'individus qui ont pris ce médicament et ont eu cet effet, ce qui n'est pas le cas sur le tableau complet car un individu peut y notifier en une seule fois plusieurs médicaments et effets.

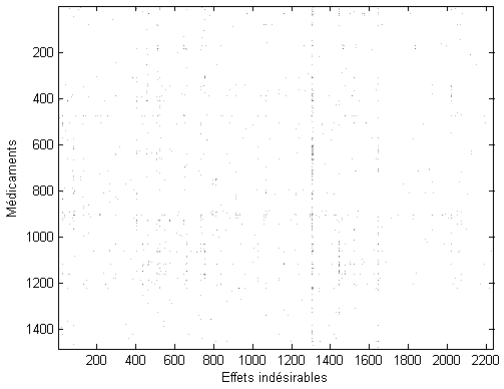


Figure 5.6 – Représentation du tableau de contingence réduit de pharmacovigilance. Le nombre de cellules non nulles est égale à 22343.

Comme nous partons du principe que les effets indésirables sont dûs à au moins un des médicaments, pour le tableau de contingence réduit, l'effet n'est dû qu'au médicament associé dans le couple. Les couples notifiés un certain nombre de fois sont donc des signaux potentiels.

De plus, parmi les 198 témoins de l'ensemble de référence $OMOP$, 76 (5 témoins négatifs et 71 témoins positifs) concernant chacun des 4 effets (mais seulement 3 témoins concernant l'effet AMI) sont présents dans le tableau de contingence réduit. Les témoins négatifs sont notifiés entre 1 et 5 fois. Les témoins positifs sont quant à eux notifiés entre 1 et 53 fois.

Nous proposons de ranger ces signaux potentiels par nombre décroissant de notifications associées et d'examiner dans cet ordre de manière pharmacologique ces signaux pour infirmer ou confirmer ces associations potentielles. Par ailleurs, il aurait été utile de connaître le nombre de personnes qui consomment un médicament donné afin de quantifier la rareté de l'effet indésirable. Les premiers signaux potentiels sont représentés dans le tableau 5.2 et notons que les témoins positifs de l'$OMOP$ ne sont pas présents dans les 50 premiers signaux que nous répertorions. Comme nous l'avons dit, ce set de référence non réalisé sur le marché français, n'est pas exhaustif et d'autres signaux potentiels autres peuvent se cacher dans la base. Nous allons discuter des premiers couples répertoriés dans le tableau 5.2, mais une analyse par des experts serait nécessaire pour analyser plus précisément les autres couples.

Discussion

Le premier couple répertorié concerne le médicament *Héparine* ($B01AB01$) et l'effet indésirable *thrombopénie* (10043554, diminution du nombre de plaquettes sanguines). Le médicament $B01AB05$ présent dans le couple 3 est également répertorié pour ce même effet et fait partie des médicaments du groupe héparine. Un rapport de la Haute Autorité de Santé datant de 2005 fait état de cette association qui a été alors étudiée plus précisément. L'association a été reconnue comme une pathologie rare et dénotée TIH (thrombopénie induite par l'héparine). Le rapport souligne que les données de pharmacovigilance sont assujetties au biais de notification, seuls les cas de thrombopénies graves avec suspicion de TIH ayant été déclarés.

Le second couple répertorié concerne le médicament *Vaccin contre l'Hépatite B* ($J07BC01$)

et l'effet indésirable *sclérose en plaques* (10048393). Un état des lieux de la causalité entre vaccination contre le virus de l'hépatite B et sclérose en plaques a été réalisé en novembre 2004. Un travail mené par une commission d'experts de l'Agence française de sécurité sanitaire des produits de santé (Afssaps), l'Agence nationale d'accréditation et d'évaluation en santé (Anaes) et l'Institut national de la santé et de la recherche médicale (Inserm). La Haute Autorité de Santé (HAS) publie le rapport d'orientation de cette commission d'audition, qui fait l'état des lieux des dernières données et études. La HAS rapporte que les données de l'étude cas-témoins de Hernan et Coll. « mettent en évidence une association entre la vaccination contre le VHB et la survenue d'une sclérose en plaques chez des adultes de 18 ans et plus ». Selon les données de la cohorte française Kidmus, la HAS rapporte : « l'absence de risque d'affection démyélinisante centrale associé à la vaccination contre le VHB chez les nouveau-nés et les nourrissons », que « les cas d'atteinte démyélinisante centrale recensés chez l'enfant et le préadolescent sont rares », et que « l'ensemble des données [...] n'exclut pas la possibilité d'un risque chez l'adulte, mais les éléments de preuve disponibles à ce jour sont insuffisants ».

Le troisième couple répertorié concerne toujours le médicament *héparine* et l'effet *embolie veineuse* (10062506). L'héparine est justement prescrite pour prévenir contre ce type d'effet et on peut s'interroger sur les raisons de la présence de ce couple dans la base réduite et qui a été notifié 356 fois (inefficacité du médicament, oubli d'une co-prescription ou d'autre effet ...). Dans ce cas, nous voyons bien l'importance d'avoir le point de vue d'un expert de la pharmacovigilance.

	Médicament (code ATC)	Effet (code Meddra)	Nombre de notifications
Couple 1	B01AB01	10043554	668
Couple 2	J07BC01	10048393	362
Couple 3	B01AB01	10062506	356
Couple 4	B01AB05	10043554	343
Couple 5	J07AN01	10022044	272
Couple 6	B01AA	10042361	252
Couple 7	B01AA	10022595	182
Couple 8	N01BB01	10013709	173
Couple 9	B01AA	GIB	169
Couple 10	A03FA01	10015832	157
Couple 11	C01BD01	10020850	156
Couple 12	M02AA10	10034972	150
Couple 13	J01CA04	10046735	148
Couple 14	J01CA04	10037844	144
Couple 15	J06BA02	10008531	138
Couple 16	B01AC06	10042361	137
Couple 17	B01AA	10018852	135
Couple 18	J06BA02	10037660	135
Couple 19	J07BB02	10022004	132
Couple 20	B01AC06	GIB	129
Couple 21	J01CA04	10002199	115
Couple 22	B01AA	10008111	114
Couple 23	J07BB02	10033775	108
Couple 24	B01AB05	10043563	108
Couple 25	N02BA01	GIB	106
Couple 26	J07BC01	10028245	105
Couple 27	J07BB02	10022086	103
Couple 28	M02AA10	10014184	99
Couple 29	J07AN01	10022095	92
Couple 30	J07BB02	10068879	92
Couple 31	N02BE01	ALI	90
Couple 32	A03FA01	10013916	90
Couple 33	C01BD01	10021114	90
Couple 34	M03AB01	10002199	80
Couple 35	C01BD01	10022611	78
Couple 36	B01AB05	10062506	78
Couple 37	J07BC01	10012305	76
Couple 38	J07AN01	10000269	74
Couple 39	B01AB06	10043554	74
Couple 40	J05AG03	10022437	71
Couple 41	M05BA08	10031264	70
Couple 42	N03AG01	10043554	70
Couple 43	J01MA12	10043255	69
Couple 44	B02BD02	10013745	67
Couple 45	B01AA	10053942	67
Couple 46	A10BA02	10023676	66
Couple 47	N02BE01	10046735	65
Couple 48	J06BA02	AKI	64
Couple 49	J05AE08	10020578	62
Couple 50	J07BB02	10033371	62

Table 5.2 – Liste des 50 premiers signaux potentiels.

5.2.1.b Classification croisée

Nous effectuons une classification croisée de ce tableau de contingence réduit, ce qui nous permettra d'obtenir une solution initiale pour la procédure *Bi-KM1* sur le tableau de contingence général. Les résultats d'une exécution de la procédure *Bi-KM1* avec la modélisation avec normalisation pour le jeu d'hyperparamètres $(4, 1, 0.01)$ sont présentés dans les tableaux 5.3 et la figure 5.7 montre le tableau de contingence réduit réorganisée à l'aide des partitions estimées. Les signaux potentiels répertoriés précédemment sont assez regroupés ensemble et sont dans des blocs avec les plus forts paramètres d'intensité γ de la loi de Poisson. En revanche, les signaux de référence de l'$OMOP$ ne sont pas généralement pas dans les mêmes blocs.

	Bi-KM1, ICL
Valeur du Critère	-577 380
$\widehat{H}$ (classe en ligne)	19
$\widehat{L}$ (classe en colonne)	20

Table 5.3 – Résultat d'une exécution de la procédure *Bi-KM1* avec normalisation pour le jeu d'hyperparamètres $(4, 1, 0.01)$ sur le tableau de contingence réduit.

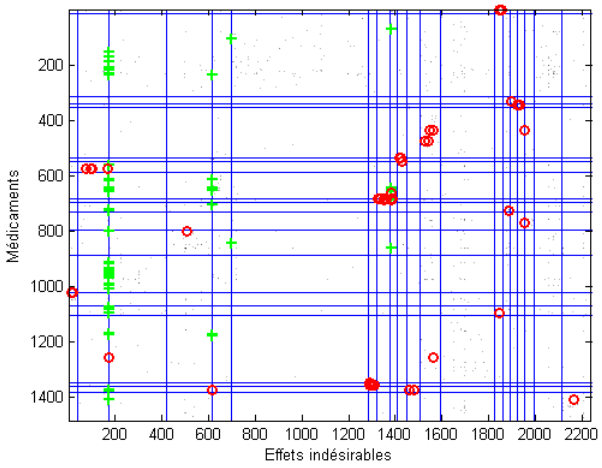


Figure 5.7 – Représentation du tableau de contingence réduit réorganisée à l'aide des partitions estimées. Les $+$ représentent les témoins positifs et les o les signaux potentiels répertoriés précédemment par la procédure.

5.2.1.c Construction de la solution initiale de la procédure BI-KM1 pour le tableau de contingence général

À partir de la solution présentée précédemment sur le tableau de contingence réduit, nous allons construire une solution initiale qui permettra d'initialiser la procédure *Bi-KM1* sur le tableau de contingence général. Afin de ne pas utiliser l'information deux fois, pour la construction du tableau de contingence général, nous écartons les 51 367 individus qui ont

servi à construire le tableau de contingence réduit. La solution proposée précédemment par la procédure *Bi-KM1* présente 19 classes en ligne et 20 classes en colonne. Mais tous les médicaments et effets indésirables ne sont pas présents dans le tableau de contingence réduit. Il faut donc affecter les médicaments et effets non présents à une des 19 et 20 classes disponibles. Nous calculons la probabilité des nouveaux médicaments et effets d'appartenir à chacune des classes à partir des paramètres estimés précédemment puis nous les affectons à la classe ayant la plus forte probabilité. Il a également fallu créer une 20 ième classe en colonne pour les médicaments qui ne présentaient aucun effet présent dans la table de contingence réduite.

5.2.2 Classification croisée du tableau de contingence

Afin de ne pas utiliser l'information deux fois, nous choisissons d'enlever les 51 367 individus qui ont servi à construire le tableau de contingence réduit, dans la construction du tableau général. À partir de la solution initiale $(20, 20)$ classes en ligne et en colonne, la procédure *Bi-KM1* a sélectionné $(40, 46)$ classes en ligne et en colonne (voir figure 5.9). Nous pouvons remarquer que la courbe d'ICL projetée sur les deux axes des classes en ligne et en colonne présente une évolution satisfaisante (voir figure 5.8).

Les premiers blocs présentant la plus forte intensité λ sont listés dans le tableau 5.4 et peuvent fournir aux pharmacologues, dans ce grand tableau de données, des premières zones intéressantes constituées d'un groupe de médicaments et d'effets réduits à étudier de manière plus précise. Nous pouvons remarquer que les signaux de l'$OMOP$ ne sont pas dans les premiers blocs listés. Les premiers blocs listés sont des blocs où les couples d'effets et de médicaments présentent peu de notifications.

Cette classification croisée du tableau de contingence constitue une information a priori très utile pour initialiser la procédure basée sur le modèle $MLBM$ (chapitre 6).

λ	n^o partition en ligne	n^o partition en colonne
0.0001	20	2
8.2640e-05	20	45
6.7897e-05	4	32
5.3217e-05	34	17
4.8757e-05	40	33
3.1252e-05	7	2
2.9201e-05	24	39
2.3247e-05	17	5
2.1353e-05	27	32
2.1108e-05	22	38

Table 5.4 – Caractéristiques des 10 premiers blocs .

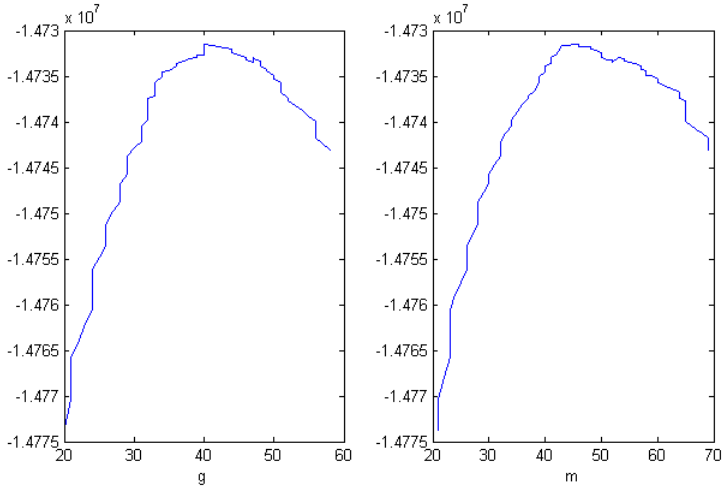


Figure 5.8 – *Courbes ICL projetées sur chaque plan (nombre de classes en ligne (à gauche) et nombre de classes en colonne (à droite)*

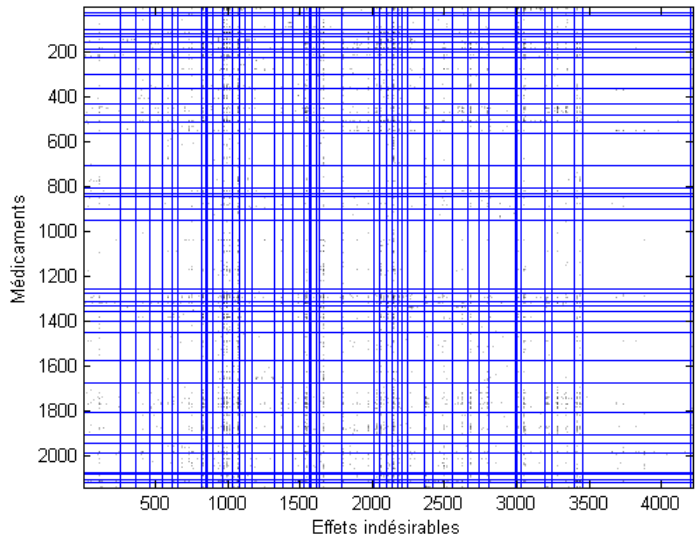


Figure 5.9 – Tableau de contingence réorganisé à l'aide des partitions estimées.

Part II

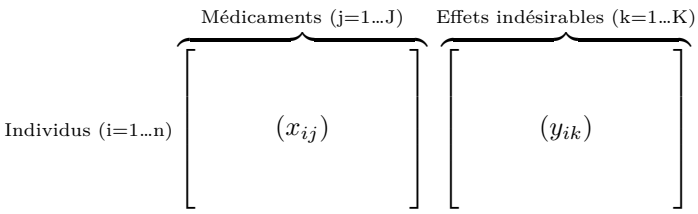
Traitement des données individuelles

6

Extension proposée : Modèle des blocs latents multiple ($MLBM$) pour les données individuelles

6.1 Définition

Les données que nous voulons traiter ici, se présentent sous la forme de deux tableaux binaires. Le premier représente la variable explicative x, réalisation d'une variable aléatoire X et le deuxième tableau, la variable réponse y, réalisation d'une variable aléatoire Y pour un certain nombre n d'individus définis par :



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

et

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

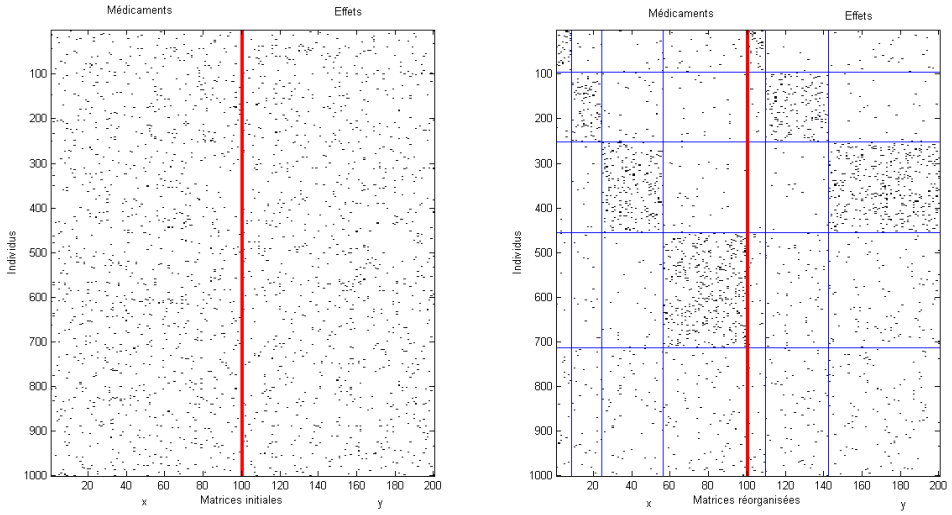


Figure 6.1 – *Matrices simulées x et y de données binaires (à gauche), réorganisée (à droite) avec la partition sur les effets, celle sur les médicaments et celle appariée sur les individus.*

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 6.1, 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.

Remarque importante. Dans ce but, nous étendons le modèle des blocs latents (Govaert and Nadif (2007)) 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. Ce modèle que nous proposons peut être vu comme un modèle des blocs latents sous contraintes.

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 $Z = (Z_1, \cdots, Z_g \cdots, Z_G)$ et en colonne $V = (V_1, \cdots, V_h \cdots, V_H)$, $W = (W_1, \cdots, W_\ell \cdots, W_L)$ sont des variables latentes.

(H_2) L'indépendance a priori d'appartenance des classes en ligne et en colonne est supposée :

$$\forall (z, v, w) \in \mathcal{Z} \times \mathcal{V} \times \mathcal{W},\ p(z, v, w) = p(z)p(v)p(w)$$

Ainsi, l'affectation en ligne et en colonne à une classe se fait de façon indépendante. Comme a priori il n'y a aucune raison de faire une différentiation entre deux lignes pour leur affectation, nous supposons que les lignes ont toutes la même probabilité d'appartenir à la $g^{\text{ème}}$ classe en ligne :

$$\forall i \in \{1, \cdots, n\},\ \mathbb{P}(Z_{ig} = 1) = \pi_g$$

Notons $\pi = (\pi_1, \cdots, \pi_G)$ 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 (resp. ℓ eme classe en colonne).
Notons $\rho = (\rho_1, \ldots, \rho_h, \ldots, \rho_H)$ et $\tau = (\tau_1, \ldots, \tau_\ell, \ldots, \tau_L)$ avec $\sum_{h=1}^{H} \rho_h = 1$ et $\sum_{\ell=1}^{L} \tau_\ell = 1$.

Avec les notations, nous obtenons :

$$p(z, v, w) = \left(\prod_{i=1}^{n} \prod_{g=1}^{G} \pi_g^{z_{ig}} \right) \left(\prod_{j=1}^{J} \prod_{h=1}^{H} \rho_h^{v_{jh}} \right) \left(\prod_{k=1}^{K} \prod_{\ell=1}^{L} \tau_\ell^{w_{k\ell}} \right).$$

(H_3) Les variables X et Y sont supposées indépendantes conditionnellement à la connaissance des appartenances en ligne et en colonne:

$$\begin{aligned} p(x, y|z, v, w) &= p(x|z, v)p(y|z, w) \\ &= \prod_{i=1}^{n} \left(\prod_{j=1}^{J} p(x_{ij}|z, v) \prod_{k=1}^{K} p(y_{ik}|z, w) \right). \end{aligned}$$

De plus, nous supposons que les variables X et Y 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)). D'où:

$$p(x, y|z, v, w) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left(\prod_{j=1}^{J} \prod_{h=1}^{H} \phi(x_{ij}; \alpha_{gh})^{z_{ig} v_{jh}} \prod_{k=1}^{K} \prod_{\ell=1}^{L} \phi(y_{ik}; \beta_{g\ell})^{z_{ig} w_{k\ell}} \right).$$

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

$$\begin{aligned} & p(x, y; \theta) \\ &= \sum_{(z,v,w) \in \mathcal{Z} \times \mathcal{V} \times \mathcal{W}} p(z; \theta)p(v; \theta)p(w; \theta)p(x|z, v; \theta)p(y|z, w; \theta) \qquad (6.1) \\ &= \sum_{(z,v,w) \in \mathcal{Z} \times \mathcal{V} \times \mathcal{W}} \prod_{i,g} \pi_g^{z_{ig}} \prod_{j,h} \rho_h^{v_{jh}} \prod_{k,\ell} \tau_l^{w_{k\ell}} \prod_{i,j,g,h} \phi(x_{ij}; \alpha_{gh})^{z_{ig} v_{jh}} \prod_{i,k,g,\ell} \phi(y_{ik}; \beta_{g\ell})^{z_{ig} w_{k\ell}}, \end{aligned}$$

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 $\theta = (\pi, \rho, \tau, \alpha, \beta)$.

6.2 Discussion autour des hypothèses

Pour la troisième hypothèse sur laquelle repose le modèle, nous avions plusieurs choix possibles :

- $p(x, y|z, v, w) = p(x|z, v, w)p(y|z, v, w)$. Mais cette hypothèse demande trop de paramètres pour le modèle. En effet, nous aurions à estimer $G-1+H-1+L-1+2\times G\times H\times L$ paramètres, où (G, H, L) représente le nombre de classes en ligne et en colonne.
- $p(x, y|z) = p(x|z)p(y|z)$. Cette hypothèse semble trop générale et pas assez spécifique à notre cas.
- $p(x, y|z, v, w) = p(x|z, v)p(y|z, w)$. Cette hypothèse est un cas particulier de l'hypothèse précédente et semble être la plus raisonnable. Même si nous supposons une indépendance entre x et y sachant les classes, x et y seront de toute manière liés par la variable latente z. De plus, nous aurons à estimer $G-1+H-1+L-1+G\times(H+L)$ paramètres, ce qui est nettement moindre que pour la première hypothèse. Nous avons choisi cette dernière hypothèse comme hypothèse (H3).

6.3 Identifiabilité

L'identifiabilité est nécessaire à la bonne estimation d'un modèle paramétrique. Rappelons qu'un modèle paramétrique est *identifiable* si la fonction $\theta \mapsto \mathbb{P}(\theta)$ est injective sur l'ensemble Θ des paramètres. Un modèle est dit *génériquement identifiable* si la fonction est injective sur Θ privé d'un ensemble de mesure nulle. Dans le cadre du modèle des blocs latents, nous avons vu dans l'introduction que Keribin et al. (2015) énonce des conditions suffisantes d'identifiabilité dans le cas de données catégorielles. En adaptant leur théorème, nous en déduisons des conditions suffisantes d'identifiabilité pour le modèle ci-dessus :

Proposition 6.3.0.1 (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 $1 \leq g \leq G$, $\pi_g > 0$ et les coordonnées des vecteurs $\kappa = A\rho$ (resp. $\omega = B\tau$) sont distinctes.*
- *C2 : Pour tout $1 \leq h \leq H$ et $1 \leq \ell \leq L$, $\rho_h > 0$ et $\tau_\ell > 0$ et les coordonnées du vecteur $\sigma = \pi' A$ (resp. $\psi = \pi' B$) sont distinctes où π' désigne la transposée de π.*
- *C3 : $n \geq \max(2H-1, 2L-1)$ et $\min(J, K) \geq 2G-1$.*

Sous ces conditions, le modèle $MLBM$ est identifiable.

Remarques.

- La condition C1 signifie que les probabilités $\mathbb{P}(x_{ij} = 1|z_{ig} = 1)$ (resp. $\mathbb{P}(y_{ik} = 1|z_{ig} = 1)$) d'obtenir un événement dans la cellule d'une ligne du tableau x (resp. du tableau y) appartenant à la classe g sont distinctes entre elles. La condition C2 concernant les colonnes des tableaux x et y s'interprète de la même manière.
- La condition C3 est naturelle et consiste à supposer qu'il y a un nombre suffisant d'observations, c'est-à-dire de lignes pour pouvoir estimer les classes en colonne.
- Sous la condition C3, le modèle $MLBM$ est *génériquement identifiable.*

6.4 Estimation des paramètres

De la même manière que le chapitre précédent, pour l'estimation des paramètres du modèle $MLBM$, nous avons recours à la même procédure proposée par Keribin et al. (2015) que nous avons adapté ici (voir section 4.2). Utilisant de manière analogue les lois conjuguées, les proportions de mélange sont munies de lois a priori de Dirichlet :

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

Nous choisissons là aussi le même hyperparamètre a pour toutes les distributions afin de ne favoriser aucune composante. En revanche, les paramètres α et β sont munis de la loi a priori Beta :

$$\alpha_{kh} \sim \beta(b, b) \quad \text{et } \beta_{k\ell} \sim \beta(b, b).$$

La modélisation bayésienne ci-dessus est résumée dans la Figure 6.2 et les algorithmes utilisés sont présentés en annexes (voir section 6.5).

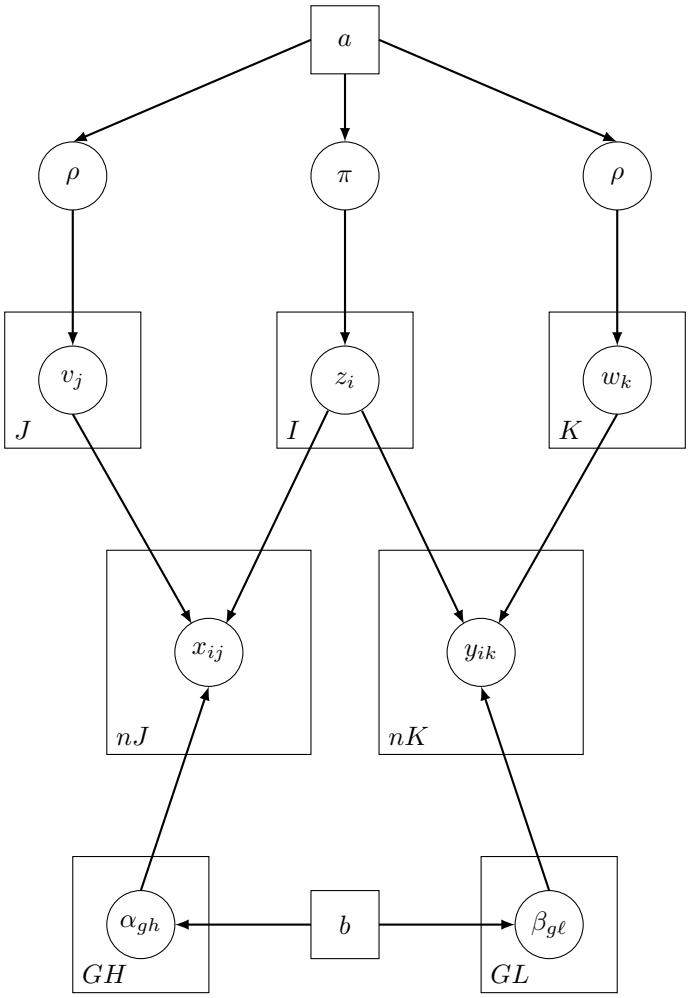


Figure 6.2 – Graphe bayésien du modèle.

6.4.1 Erreur de classification

Afin de comparer des partitions entre elles, nous choisissons de prendre l'erreur de classification proposée par Lomet et al. (2012) qui semble pour l'instant plus facile à adapter que celle induite par le *Coclustering Adjusted Index*. Pour cela nous introduisons la distance de classification suivante sur les lignes entre deux partitions z et z' par :

$$\text{dist}_{n,G}(z, z') = 1 - \max_{\sigma \in \mathfrak{S}(\{1,..G\})} \frac{1}{n} \sum_{i,g} z_{ig} z'_{i\sigma(g)}$$

De manière symétrique, nous définissons la distance de classification sur les colonnes de x notée $\text{dist}_{J,H}$, et sur celles de y, notée $\text{dist}_{K,L}$.

La distance de classification entre deux triplets (z, v, w) et (z', v', w') que nous proposons est la suivante :

$$\begin{aligned} & \text{dist}_{(n,G)\times(J,H)\times(K,L)}((z, v, w); (z', v', w')) \\ = \; & 1 - \max_{\sigma} \max_{\tau} \max_{\eta} \frac{1}{n(J+K)} \sum_{i,g} z_{ig} z'_{i\sigma(g)} \left(\sum_{j,h} v_{jh} v'_{j\tau(h)} + \sum_{k,\ell} w_{k\ell} w'_{k\eta(\ell)}) \right), \end{aligned}$$

où $\sigma \in \mathfrak{S}(\{1,..G\}), \tau \in \mathfrak{S}(\{1,..H\})$ et $\eta \in \mathfrak{S}(\{1,..L\})$. Par ailleurs, la distance précédente peut s'écrire également si on note vw la concaténation de v et de w:

$$\begin{aligned} \text{dist}_{(n,G)\times(J+K,H+L)}((z, vw); (z', (vw)')) \; &= \; \text{dist}_{n,g}(z; z') + \text{dist}_{J+K,H+L}(vw; (vw)') \\ &- \; \text{dist}_{n,G}(z; z') \times \text{dist}_{J+K,G+L}(vw; (vw)') \end{aligned}$$

Remarque. Nous parlons d'erreur de classification d'un triplet (z, v, w) lorsque nous calculons la distance de classification de celui-ci avec les vraies partitions (z^*, v^*, w^*) :

$$e_{(n,G)\times(J,H)\times(K,L)}(z, v, w) = \text{dist}_{(n,G)\times(J,H)\times(K,L)}((z, v, w); (z^*, v^*, w^*)$$

6.4.2 Discussion autour du choix des hyperparamètres

Keribin et al. (2015) prônent le choix suivant pour les hyperparamètres $(a, b) = (4, 1)$ dans le cas du LBM binaire. Nous suivons leur préconisation en ce qui concerne le choix de a qui a un effet bénéfique contre la dégénérescence des classes. Remarquons que choisir $b = 1$ revient à prendre des α suivant une loi uniforme alors que les α pour les matrices creuses sont généralement très petits :

$$\forall (g, h) \in \{1, \cdots, G\} \times \{1, \cdots, H\}, \quad \alpha_{gh} \sim \mathcal{B}e(1, 1).$$

Pour tester et valider ce choix d'hyperparamètres, Nous simulons des matrices avec un nombre de lignes et de colonnes (n, J, K) suivant : $(1000, 100, 100)$ et à classes fixées (G, H, L)=(5,4,3). De plus, nous proposons un plan d'expérience inspiré du protocole de Lomet et al. (2012) qui permettent de générer des matrices plus ou moins faciles à classifier, calibrées selon l'erreur de classification introduite précédemment (5%, 15% et 25% d'erreur codés par $+$, $++$ et $+++$). Nous étudierons également le cas de proportions de mélanges équilibrés ou déséquilibrés.

Par ailleurs, les matrices simulées vérifient largement les critères d'identifiabilité, soit que la distance minimale entre deux cordonnées des vecteurs κ, ω, σ et ψ est d'au moins 10^{-3}. De plus, elles ne comportent aucune ligne ni aucune colonne nulle.

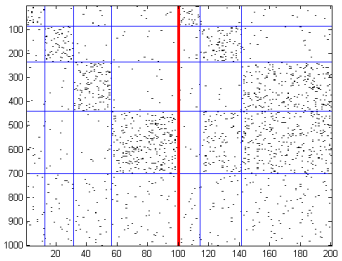


Figure 6.3 – Exemple de matrices pour des proportions avec une progression arihtmétique et dans le cas difficile.

Un exemple de telle matrice est fournie par la figure 6.3. Enfin, nous comparons différentes stratégies, l'initialisation aléatoire couplé avec l'algorithme *V-Bayes*, l'initialisation *small V-Bayes* (plusieurs lancements de l'algorithme *V-Bayes* avec peu d'itérations) couplé avec l'algorithme *V-Bayes*, et l'échantillonneur de Gibbs couplé avec l'algorithme *V-Bayes* et l'échantillonneur de Gibbs seul. Ces stratégies sont comparées à temps fixé, avec comme référence, le temps mis par l'échantillonneur de Gibbs pour 200 000 itérations.

6.4.2.a Proportions équilibrées

Dans ce cadre, nous simulons :

- 50 matrices pour chaque difficulté modélisée par ϵ, telles que

$$\alpha = \begin{pmatrix} 5\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 10\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 15\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 20\epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon \end{pmatrix} \text{ et } \beta = \begin{pmatrix} 5\epsilon & \epsilon & \epsilon \\ \epsilon & 10\epsilon & \epsilon \\ \epsilon & \epsilon & 15\epsilon \\ \epsilon & 10\epsilon & 20\epsilon \\ \epsilon & \epsilon & \epsilon \end{pmatrix},$$

  avec ϵ variant de 0 à 0.5 $\longrightarrow$ matrices ($+$ à $+++$) et,

- avec proportions équilibrées : $\pi = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$, $\rho = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$ et $\tau = \begin{pmatrix} 0.33 \\ 0.33 \\ 0.34 \end{pmatrix}$.

Les résultats sont présentés dans la figure 6.4. Nous remarquons que les stratégies de l'échantillonneur de Gibbs couplé avec l'algorithme *V-Bayes* et l'échantillonneur de Gibbs seul sont nettement moins performantes que les deux premières lorsque les données sont très mélangées. En analysant les résultats plus précisément, nous nous rendons compte que l'échantillonneur de Gibbs fournit des solutions avec des classes vides et ne parvient pas à les reremplir dans un temps fini raisonnable.

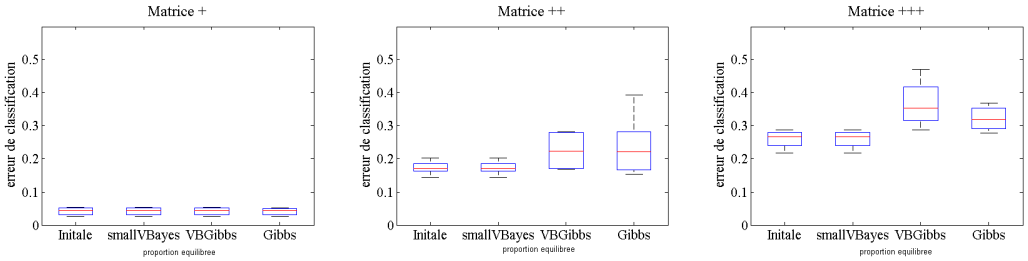


Figure 6.4 – Comparaison des boxplots des erreurs de classification des différentes stratégies utilisées en fonction de la difficulté des matrices (celle-ci augmente de gauche vers la droite).

6.4.2.b Proportions déséquilibrées

Dans ce cadre, nous simulons :

- 50 matrices pour chaque difficulté modélisée par ϵ,

  avec
  $$\alpha = \begin{pmatrix} 10\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 11\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 12\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 13\epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon \end{pmatrix} \text{ et } \beta = \begin{pmatrix} 10\epsilon & \epsilon & \epsilon \\ \epsilon & 11\epsilon & \epsilon \\ \epsilon & \epsilon & 12\epsilon \\ \epsilon & 10\epsilon & 11\epsilon \\ \epsilon & \epsilon & \epsilon \end{pmatrix},$$

  avec ϵ variant de 0 à 0.5 $\longrightarrow$ matrices ($+$ à $+++$),

- et avec proportions suivant une progression arithmétique :
  $$\pi = \begin{pmatrix} 0.1 \\ 0.15 \\ 0.2 \\ 0.25 \\ 0.3 \end{pmatrix}, \ \rho = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{pmatrix} \text{ et } \tau = \begin{pmatrix} 0.1 \\ 0.33 \\ 0.57 \end{pmatrix}.$$

Les résultats sont présentés dans la figure 8.8 et sont sensiblement les mêmes que ceux relatifs aux proportions équilibrées.

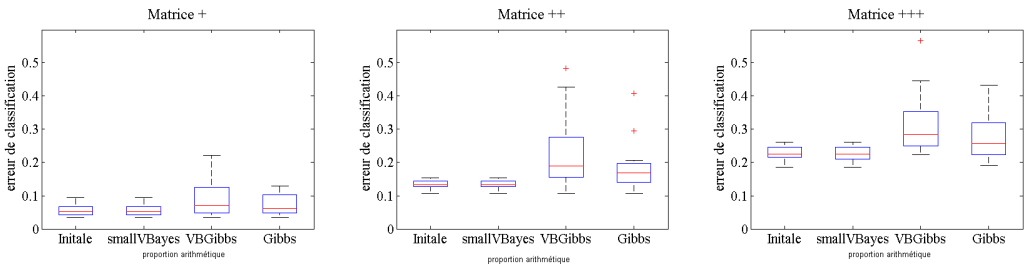


Figure 6.5 – Comparaison des boxplots des erreurs de classification des différentes stratégies utilisées en fonction de la difficulté des matrices (celle-ci augmente de gauche vers la droite).

6.4.2.c Discussion

Dans le cas du $MLBM$ binaire et de données très creuses, l'échantillonneur de Gibbs semble être sensible au choix de la loi a priori des α. Montrons alors que pour l'échantillonneur de Gibbs, la probabilité de reremplir une classe vide est très petite dans le cas d' une matrice creuse et des lois a priori précédentes pour les α.

Dans un premier temps, nous nous plaçons dans le cadre du LBM. Supposons que $G = 2$ classes en lignes et $H = 1$ classe en colonne et fixons le nombre (n, J) de lignes et de colonnes pour x à (1000,100).

De plus, supposons qu'à l'étape (d) de l'algorithme, la classe 1 en ligne est vide et la classe 2 avec une estimation de α_2 de l'ordre du degré de sparsité de la matrice, par exemple 0.01.

Regardons un individu i de la matrice qui globalement est au moins de cette forme :

1	0	$\cdots$	$\cdots$	0

Ainsi, à l'étape $(d+1)$, la probabilité conditionnelle que l'indidividu i appartienne à la classe vide 1, vaut :

$$\begin{aligned}
\mathbb{P}(Z_{i1} = 1|x, v, \theta) &= \frac{\pi_1 \prod_j \alpha_1^{x_{ij}}(1-\alpha_1)^{1-x_{ij}}}{\pi_1 \prod_j \alpha_1^{x_{ij}}(1-\alpha_1)^{1-x_{ij}} + \pi_2 \prod_j \alpha_2^{x_{ij}}(1-\alpha_2)^{1-x_{ij}}} \\
&= \frac{\pi_1 \alpha_1^1 (1-\alpha_1)^{99}}{\pi_1 \alpha_1^1 (1-\alpha_1)^{99} + \pi_2 \alpha_2^1 (1-\alpha_2)^{99}} \\
&= \frac{1}{1 + \frac{\pi_2}{\pi_1}\frac{\alpha_2}{\alpha_1}(\frac{1-\alpha_2}{1-\alpha_1})^{99}}.
\end{aligned}$$

Or $\alpha_2 \approx 0.01$, $\alpha_1 \sim \mathcal{B}e(1,1)$, donc ce dernier a $\frac{9}{10}$ d'être supérieur à 0.1. Prenons donc un des cas les plus critiques à savoir $\alpha_1 = 0.1$. On a donc

$$\frac{\alpha_2}{\alpha_1} = 0.1, \quad \frac{1-\alpha_2}{1-\alpha_1} = 1.1.$$

Ensuite, comme $\pi \sim \mathcal{D}(a + z_{+1}, \cdots, a + z_{+G})$, $E(\pi_1) = \frac{4+0}{2\times 4+1000} = \frac{4}{1008}$ et $E(\pi_2) = \frac{1004}{1008}$.

Ainsi, nous obtenons finalement,

$$P(Z_{i1} = 1|x, v, \theta) \approx 3 \times 10^{-6}.$$

Remarque. Si $\alpha_1 = 0.5$, on a $P(Z_{i1} = 1|x, v, \theta) \approx 8 \times 10^{-31}$.

En augmentant le nombre de classes en colonnes et en ligne et en considérant cette fois-ci les deux tableaux, nous voyons que la probabilité estimée va être encore plus petite que la valeur calculée précédemment dans le cas du LBM.

Nous concluons que pour l'échantillonneur de Gibbs, la loi a priori $\mathcal{B}e(1,1)$ est bien adaptée dans le cas de matrices non creuses (Keribin et al. (2015)), mais qu'il vaudrait mieux adapter la loi a priori en fonction de la sparsité de la matrice.

Ainsi, nous choisissons la loi a priori suivante, $\mathcal{B}e(2, 100)$ pour l'échantillonneur de Gibbs, dont la densité présente un fort a priori autour de 0.02 et est représentée par la figure 6.6. L'algorithme *V-Bayes* reste utilisé avec le choix d'hyperparamètres $(a, b) = (4, 1)$.

Nous choisissons de reproduire le plan d'expérience précédent en simulant des matrices où le nombre moyen de cases non nulles est de l'ordre de cette valeur. Les résultats (voir figures 8.8 et 6.7) montrent que les différentes stratégies présentent des performances similaires. Toutefois, nous notons la présence d'outliers chez les deux dernières stratégies. Le choix de cette loi a priori semble être adéquat pour ce type de données simulées. Ainsi, il faudrait

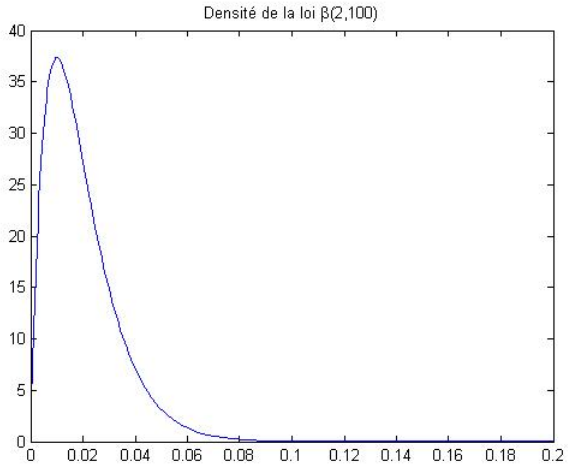


Figure 6.6 – Densité de la loi $\mathcal{B}e(2, 100)$

adapter les coefficients b_1 et b_2 de la loi $\mathcal{B}e(b_1, b_2)$ en fonction de la sparsité de la matrice étudiée. En effet, le pic de cette densité devrait se situer aux environs de la moyenne empirique des cases non nulles de la matrice, tout en essayant d'avoir la plus grande variance possible pour rester la moins informative possible (ce qui semble difficile car cette loi ne semble pas très flexible). Une étude plus poussée serait à réaliser et constitue une perspective intéressante de ce travail. Au vu de ces constats, nous avons choisi d'utiliser la stratégie *small V-Bayes* couplé avec l'algorithme *V-Bayes* qui semble plus cohérente avec le même choix d'hyperparamètres $(4, 1)$. Ici, la comparaison des stratégies a été effectuée à temps fixé mais le nombre de lancements de la procédure *small V-Bayes*, similaire à la procédure *small EM* reste une question classique. Nous choisissons d'en effectuer 200 dans tout ce qui suit.

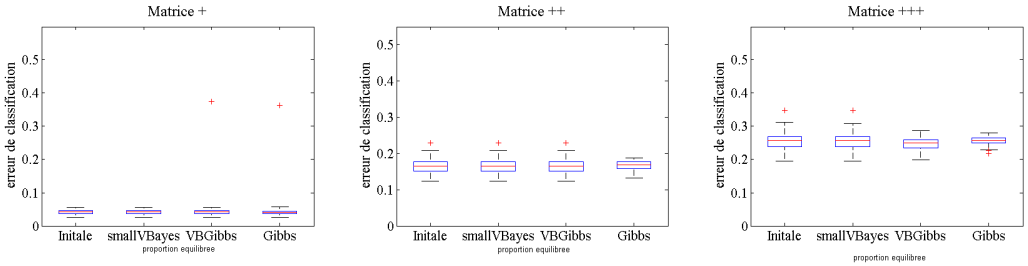


Figure 6.7 – Comparaison des boxplots des erreurs de classification des différentes stratégies utilisées en fonction de la difficulté des matrices (celle-ci augmente de gauche vers la droite) pour proportions équilibrées.

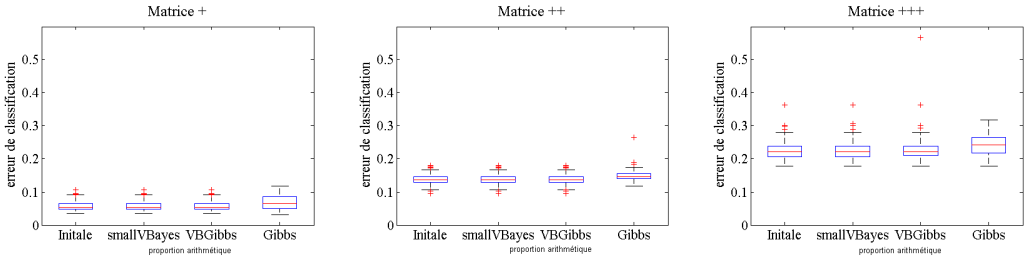


Figure 6.8 – Comparaison des boxplots des erreurs de classification des différentes stratégies utilisées en fonction de la difficulté des matrices (celle-ci augmente de gauche vers la droite) pour proportions déséquilibrées.

6.5 Annexes

6.5.0.a Formulaire de l'échantillonneur de Gibbs

Le schéma numérique de l'échantillonneur de Gibbs présenté dans la section 4.2.3 et adapté dans le cadre du modèle $MLBM$ est le suivant :

Échantillonneur de Gibbs.

Obtention d'une chaîne de Markov de loi stationnaire $p(z, v, w, \theta|x, y)$.

1. Initialisation de $\theta^{(0)}$, de $v^{(0)}$ et de $w^{(0)}$.
2. Pour $d = 0 \dots n_{\text{iter}}$:
   - Simulation de $z^{(d+1)}$ suivant la loi $p(z|x, y, v^{(d)}, w^{(d)}; \theta^{(d)})$.
   - Simulation de $v^{(d+1)}$ suivant la loi $p(v|x, y, z^{(d+1)}; \theta^{(d)})$.
   - Simulation de $w^{(d+1)}$ suivant la loi $p(w|x, y, z^{(d+1)}, v^{(d+1)})$.
   - Simulation de $\theta^{(d+1)}$ suivant la loi $p(\theta|x, y, z^{(d+1)}, v^{(d+1)}, w^{(d+1)})$.
3. Obtention d'un estimateur $\widehat{\theta}^G = \frac{1}{n_{iter}} \sum_{c=1}^{n_{iter}} \theta^{(d)}$.

Nous pouvons expliciter chacune des lois intervenant dans l'algorithme. Pour cette partie, nous avons encore une symétrie entre z, v, et w et nous ferons que dans le cas de z :

$$\begin{aligned}
\mathbb{P}(z_{ig} = 1|v, w, x, y, \theta) &= \frac{\mathbb{P}(z_{ig} = 1|\theta)\mathbb{P}(x_{i.}, y_{i.}|z_{ig} = 1, x, y, v, w, \theta)}{\mathbb{P}(x_{i.}, y_{i.}|v, w, \theta)} \\
&= \frac{\pi_g \prod_j \left[\mathbb{P}(x_{ij}|z_{ig} = 1, x, y, v, w, \theta)\right] \prod_k \left[\mathbb{P}(y_{ik}|z_{ig} = 1, x, y, v, w, \theta)\right]}{\sum_{g'} \left(\pi_{g'} \prod_j \left[\mathbb{P}(x_{ij}|z_{ig'} = 1, x, y, v, w, \theta)\right] \prod_k \left[\mathbb{P}(y_{ik}|z_{ig'} = 1, x, y, v, w, \theta)\right)\right.} \\
&= \frac{\pi_g \prod_{j,h} \left[\left(\frac{\alpha_{gh}}{1-\alpha_{gh}}\right)^{v_{jh}x_{ij}} (1-\alpha_{gh})^{v_{jh}}\right] \prod_{k,\ell} \left[\left(\frac{\beta_{g\ell}}{1-\beta_{g\ell}}\right)^{w_{k\ell}y_{ik}} (1-\beta_{g\ell})^{w_{k\ell}}\right]}{\sum_{g'} \left(\pi_{g'} \prod_{j,h} \left[(\alpha_{g'h})^{v_{jh}x_{ij}} (1-\alpha_{g'h})^{v_{jh}(1-x_{ij})}\right] \prod_{k,\ell} \left[(\beta_{g'\ell})^{w_{k\ell}y_{ik}} (1-\beta_{g'\ell})^{w_{k\ell}(1-y_{ik})}\right]\right)}
\end{aligned}$$

Par symétrie :

$$\mathbb{P}(v_{jh}=1|v,w,x,y,\theta) \quad = \quad \frac{\rho_h \prod_{i,g}\left[\left(\frac{\alpha_{gh}}{1-\alpha_{gh}}\right)^{z_{ig}x_{ij}}(1-\alpha_{gh})^{z_{ig}}\right]}{\sum_{h'}\left(\rho_{h'}\prod_{i,g}\left[(\alpha_{gh'})^{z_{ig}x_{ij}}(1-\alpha_{gh'})^{z_{ig}(1-x_{ij})}\right]\right)}.$$

On a de même une formule similaire pour $\mathbb{P}(w_{k\ell}=1|v,w,x,y,\theta)$. De même, il y a une symétrie entre π et ρ, τ. Nous avons avec π la loi a priori :

$$\begin{aligned}
\boldsymbol{\pi}(\pi|z,v,w,\rho_h,\tau_\ell,\alpha_{gh},\beta_{g\ell},x,y) &\propto \mathbb{P}(z|\pi,v,w,\rho_h,\tau_\ell,\alpha_{gh},\beta_{g\ell},x,y)\boldsymbol{\pi}(\pi|v,w,\rho_h,\tau_\ell,\alpha_{gh},\tau_{g\ell},x,y)\\
&\propto \mathbb{P}(z|\pi)\boldsymbol{\pi}(\pi)\\
&\propto \prod_{i,g}\pi_g^{z_{ig}}\prod_g \pi_g^{a-1}\\
&\propto \prod_g \pi_g^{(a+\sum_i z_{ig})-1)}\\
&\propto \prod_g \pi_g^{(a+z_{+g})-1)}
\end{aligned}$$

en notant $z_{+g}=\sum_i z_{ig}$. Donc nous avons

$$\pi|z,v,w,\rho_h,\tau_\ell,\alpha_{gh},\beta_{g\ell},x,y \sim \mathcal{D}(a+z_{+1},\cdots,a+z_{+G}).$$

De même, en notant $v_{+h}=\sum_j v_{jh}$ et $w_{+\ell}=\sum_k w_{k\ell}$, on a

$$\rho|z,v,w,\pi,\tau_\ell,\alpha_{gh},\beta_{g\ell},x,y \sim \mathcal{D}(a+v_{+1},\cdots,a+v_{+H})$$

et

$$\tau|z,v,w,\rho_h,\pi,\alpha_{gh},\beta_{g\ell},x,y \sim \mathcal{D}(a+w_{+1},\cdots,a+w_{+L}).$$

Il reste à calculer la loi de $\alpha_{gh}|x,z,v$. Le calcul sera similaire pour la loi de $\beta_{g\ell}|y,z,w$. Nous avons :

$$\begin{aligned}
\boldsymbol{\pi}(\alpha_{gh}|x,z,v) &\propto \mathbb{P}(x|z,v,\alpha_{gh})\prod_{g,h}\boldsymbol{\pi}(\alpha_{gh})\\
&\propto \prod_{i,j,g,h}\left[\phi(x_{ij};\alpha_{gh})^{z_{ig}v_{jh}}\left(\alpha_{gh}(1-\alpha_{gh})\right)^{b-1}\right]\\
&\propto \prod_{g,h}\left[\alpha_{gh}^{\left(\sum_{i,j}x_{ij}z_{ig}v_{jh}+b\right)-1}\left(1-\alpha_{gh}\right)^{\left(\sum_{i,j}z_{ig}v_{jh}-\sum_{i,j}x_{ij}z_{ig}v_{jh}+b\right)-1}\right].
\end{aligned}$$

Nous voyons que les paramètres pour chaque bloc sont indépendants et nous avons :

$$\alpha_{gh}|x,z,v \sim \mathcal{B}e(\delta_{gh},\eta_{g\ell}-\delta_{g\ell}+2b),$$

avec $\delta_{gh}=\sum_{i,j}x_{ij}z_{ig}v_{jh}+b$ et $\eta_{g\ell}=\sum_{i,j}z_{ig}w_{j\ell}$.

6.5.0.b Formulaire de l'algorithme V-Bayes

Nous obtenons l'adaptation de l'algorithme *V-Bayes* cherchant toujours à maximiser l'énergie libre $\mathcal{F}_B$ définie par :

$$\mathcal{F}_B(\theta) = \mathcal{F}(\theta) + \log p(\theta).$$

Algorithme *V-Bayes* pour le $MLBM$.

1. Initialisation de $\theta^{(0)}$, de $r_{jh}^{(0)}$ et de $t_{k\ell}^{(0)}$.

2. Pour $d = 0 \dots n_{\text{iter}}$:

   - Étape VE : maximisation alternée de l'énergie libre à $\theta^{(d)}$ fixé en prenant $r_{jh}^{(t=0)} = r_{jh}^{(d)}$ et $t_{k\ell}^{(t=0)} = t_{k\ell}^{(d)}$:

     - calcul de $s_{ig}^{(t+1)}$ à $r_{jh}^{(t)}$, à $t_{k\ell}^{(t)}$ et à $\theta^{(d)}$ fixés:

     $$s_{ig}^{(t+1)} = \frac{\pi_g^{(d)} f_{ig}\left(\alpha_{gh}^{(t+1)}, r_{jh}^{(t+1)}, x_{ij}\right) f_{ig}\left(\beta_{g\ell}^{(t+1)}, t_{k\ell}^{(t+1)}, y_{ik}\right)}{\sum_{g'=1}^{G}\left(\pi_{g'}^{(d)} f_{ig'}\left(\alpha_{g'h}^{(t+1)}, r_{jh}^{(t+1)}, x_{ij}\right) f_{ig'}\left(\beta_{g'\ell}^{(t+1)}, t_{k\ell}^{(t+1)}, y_{ik}\right)\right)},$$

     avec $f_{ig}(\alpha_{gh}, r_{jh}, x_{ij}) = \prod_{h=1}^{H}\left[\alpha_{gh}^{\sum_{j=1}^{J} r_{jh}x_{ij}} (1-\alpha_{gh})^{\sum_{j=1}^{J} r_{jh}(1-x_{ij})}\right]$.

     - calcul de $r_{jh}^{(t+1)}$ et $t_{k\ell}^{(t+1)}$ à $s_{ig}^{(t+1)}$ et à $\theta^{(d)}$ fixés.

   $\longrightarrow$ Obtention des probabilités $s_{ig}^{(d+1)}$, $r_{jh}^{(d+1)}$ et $t_{k\ell}^{(d+1)}$.

   - Étape M : calcul du paramètre $\theta^{(d+1)}$:

   $$\pi_g^{(d+1)} = \frac{a - 1 + \sum_{i=1}^{n} s_{ig}^{(d+1)}}{G(a-1) + n}, \quad \rho_h^{(d+1)} = \frac{a - 1 + \sum_{j=1}^{J} r_{jh}^{(d+1)}}{H(a-1) + J}, \quad \tau_l^{(d+1)} = \frac{a - 1 + \sum_{k=1}^{K} t_{k\ell}^{(d+1)}}{L(a-1) + K}$$

   $$\text{et } \alpha_{gh}^{(d+1)} = \frac{b - 1 + \sum_{i=1}^{n}\sum_{j=1}^{J} s_{ig}^{(d+1)} r_{jh}^{(d+1)} x_{ij}}{2(b-1) + \sum_{i=1}^{n} r_{jh}^{(d+1)} x_{ij}}, \quad \beta_{g\ell}^{(d+1)} = \frac{b - 1 + \sum_{i=1}^{n}\sum_{k=1}^{K} s_{ig}^{(d+1)} t_{k\ell}^{(d+1)} y_{ik}}{2(b-1) + \sum_{i=1}^{n} t_{k\ell}^{(d+1)} y_{ik}}.$$

3. Obtention d'un estimateur $\widehat{\theta}^{MVB} = \theta^{(n_{\text{iter}})}$.

$\longrightarrow$ Calcul des estimateurs des partitions (MAP) :

$$(\widehat{z}_i)^{MVB} \in \underset{g=1,\dots,G}{\text{argmax}}\, s_{ig}^{(n_{\text{iter}})}, (\widehat{v}_j)^{MVB} \in \underset{h=1,\dots,H}{\text{argmax}}\, r_{jh}^{(n_{\text{iter}})}$$
$$\text{et } (\widehat{w}_k)^{MVB} \in \underset{\ell=1,\dots,L}{\text{argmax}}\, t_{k\ell}^{(n_{\text{iter}})}$$

Plus précisément la fonction à maximiser est la suivante

$$
\begin{aligned}
\mathbb{E}_{q_{vw}}\left[\log p(c,v,w|\theta)\right] + \log p(\theta) \quad = \quad & \sum_{j,h} r_{jh} \log \rho_h + \sum_{k,\ell} t_{k\ell} \log \tau_\ell \\
& + \log \frac{\Gamma(Ha)}{\Gamma(a)^H} + (a-1) \sum_h \log \rho_h + \log \frac{\Gamma(La)}{\Gamma(a)^L} + (a-1) \sum_\ell \log \tau_\ell \\
& + \sum_{j,k,h,\ell} r_{jh} t_{k\ell} \left[-\lambda_{k\ell} - \mu_j - \nu_k + c_{jk} (\log \lambda_{k\ell} + \log \mu_j + \log \nu_k) \right] \\
& - \sum_{h,\ell} r_{jh} t_{k\ell} \sum_{j,k} \log c_{jk}! \\
& + gm \left(\alpha \log \beta - \log \Gamma(\alpha) \right) + (\alpha - 1) \sum_{k,\ell} \log \lambda_{k\ell} - \beta \sum_{k,\ell} \lambda_{k\ell}
\end{aligned}
$$

Les parties en rouge étant constantes, nous ne sommes pas obligés de les inclure.

Ainsi, cela revient à maximiser la fonction suivante, les parties en rouge étant des constantes et en remplaçant ϕ et en notant $u_{ih} = \sum_j r_{jh} x_{ij}$ et $d_h = \sum_j r_{jh}$ (de même pour $\mu_{i\ell}$ et ν_ℓ):

$$
\begin{aligned}
g(s,r,t,\theta) \quad = \quad & \sum_{i,g} s_{ig} \log \left[\pi_g \prod_l \left(\frac{\alpha_{gh}}{1-\alpha_{gh}}\right)^{u_{ih}} (1-\alpha_{gh})^{d_h} \prod_\ell \left(\frac{\beta_{g\ell}}{1-\beta_{g\ell}}\right)^{\mu_{i\ell}} (1-\beta_{g\ell})^{\nu_\ell} \right] \\
& + \sum_{j,h} r_{jh} \log \rho_h + \sum_{k,\ell} t_{k\ell} \log \tau_\ell \\
& - \sum_{i,g} s_{ig} \log s_{ig} - \sum_{j,h} r_{jh} \log r_{jh} - \sum_{k,\ell} t_{k\ell} \log t_{k\ell} \\
& + (a-1) \sum_g \log \pi_g + (a-1) \sum_h \log \rho_h + (a-1) \sum_\ell \log \tau_\ell \\
& + \sum_g \left(\sum_h (b-1) \log \left[\alpha_{gh}(1-\alpha_{gh})\right] + \sum_\ell (b-1) \log \left[\beta_{g\ell}(1-\beta_{g\ell})\right] \right).
\end{aligned}
$$

Pour satisfaire en plus les conditions $\sum_g s_{ig} = 1$, $\sum_h r_{jh} = 1$ et $\sum_\ell t_{j\ell} = 1$, on utilise le lagrangien et nous voulons donc maximiser la fonction suivante :

$$
h(s,r,t,\theta,\lambda,\zeta,\gamma) = g(s,r,t,\theta) + \sum_i \lambda_i \left(\sum_g s_{ig} - 1 \right) + \sum_h \zeta_h \left(\sum_l r_{jh} - 1 \right) + \sum_\ell \gamma_\ell \left(\sum_\ell t_{j\ell} - 1 \right).
$$

En dérivant nous obtenons :

$$
\begin{aligned}
\frac{\partial}{\partial s_{ig}} h(s,t,\theta,\lambda,\zeta,\gamma) = 0 \quad & \Leftrightarrow \quad \log \left[\pi_g \prod_h \left(\frac{\alpha_{gh}}{1-\alpha_{gh}}\right)^{u_{ih}} (1-\alpha_{gh})^{d_h} \prod_\ell \left(\frac{\beta_{g\ell}}{1-\beta_{g\ell}}\right)^{\mu_{i\ell}} (1-\beta_{g\ell})^{\nu_\ell} \right] - \log s_{ig} - 1 + \lambda_i = 0 \\
& \Leftrightarrow \quad s_{ig} = e^{\lambda_i - 1} \pi_g \prod_h \left[\left(\frac{\alpha_{gh}}{1-\alpha_{gh}}\right)^{u_{ih}} (1-\alpha_{gh})^{d_h} \right] \prod_\ell \left[\left(\frac{\beta_{g\ell}}{1-\beta_{g\ell}}\right)^{\mu_{i\ell}} (1-\beta_{g\ell})^{\nu_\ell} \right].
\end{aligned}
$$

En dérivant sur λ on obtient que $\sum_g s_{ig} = 1$ ce qui donne finalement la formule de s_{ig} dans l'algorithme *V-Bayes*.

Remarque. En pratique, on calculera $\log s_{ig}$.
On obtient de même une formule similaire pour chacun des r_{jh} et $t_{k\ell}$.

Pour l'étape M, nous allons chercher à maximiser

$$f(s,r,,\theta,\lambda,\mu,\gamma) = g(s,r,t,\theta) + \lambda\Big(\sum_g \pi_g - 1\Big) + \mu\Big(\sum_h \rho_{jh} - 1\Big) + \gamma\Big(\sum_\ell \tau_{j\ell} - 1\Big).$$

La maximisation étant la même pour π, ρ et τ, nous ferons uniquement dans ce document celle sur π_g:

$$\begin{aligned}\frac{\partial}{\partial \pi_g} f(s,r,t,\theta,\lambda,\mu,\gamma) = 0 \quad &\Leftrightarrow \quad \frac{1}{\pi_g}\sum_i s_{ig} + \frac{a-1}{\pi_g} + \lambda = 0 \\ &\Leftrightarrow \quad \pi_g = \frac{a-1+\sum_i s_{ig}}{-\lambda}.\end{aligned}$$

En dérivant par rapport à λ, on a donc que $\sum_g \pi_g = 1$ soit :

$$\pi_g = \frac{a-1+\sum_i s_{ig}}{G(a-1)+n}.$$

De même nous avons :

$$\rho_h = \frac{a-1+\sum_j r_{jh}}{H(a-1)+J}, \text{ et } \tau_\ell = \frac{a-1+\sum_k t_{k\ell}}{L(a-1)+K}.$$

Concernant la maximisation de la fonction par rapport à α_{gh}, nous obtenons :

$$\begin{aligned}\frac{\partial}{\partial \alpha_{gh}} f(s,r,t,\theta,\lambda,\mu,\gamma) = 0 \quad &\Leftrightarrow \quad \sum_{i,j} s_{ig} r_{jh} x_{ij} \frac{1}{\alpha_{gh}} - \sum_{i,j} s_{ig} r_{jh}(1-x_{ij}) \frac{1}{1-\alpha_{gh}} + \frac{(b-1)}{\alpha_{gh}} - \frac{(b-1)}{1-\alpha_{gh}} = 0 \\ &\Leftrightarrow \quad \sum_{i,j} s_{ig} r_{jh} x_{ij}(1-\alpha_{gh}) - \sum_{i,j} s_{ig} r_{jh}(1-x_{ij})\alpha_{gh} + (b-1)(1-\alpha_{gh}) - (b-1)\alpha_{gh} = 0 \\ &\Leftrightarrow \quad \alpha_{gh}\Big(\sum_{i,j} s_{ig} r_{jh} x_{ij} + \sum_{i,j} s_{ig} r_{jh} - \sum_{i,j} s_{ig} r_{jh} x_{ij} + 2(b-1)\Big) \\ &- \quad (b-1) - \sum_{i,j} s_{ig} r_{jh} x_{ij} = 0.\end{aligned}$$

7

Sélection de modèles pour le $MLBM$

Nous nous intéressons ici à l'adaptation du critère d'inspiration bayésienne ICL (*Integrated Completed Likelihood*) pour le $MLBM$. Nous proposons également, à la manière de Keribin et al. (2015), une forme pour le critère BIC pour le modèle étudié.

7.1 Sélection de modèles pour le modèle $MLBM$

7.1.1 Critère Intregated Completed Likelihood (ICL)

Le principe du critère $ICL(G, H, L)$ pour le modèle $MLBM$ cherche toujours à maximiser la log-vraisemblance complétée $\log p(x, y, z, v, w|G, H, L)$. Calculons $p(x, y, z, v, w|G, H, L)$ en utilisant l'indépendance conditionnelle par rapport à θ de v et w et l'hypothèse (H3) du modèle :

$$
\begin{aligned}
& \int p(x,y,z,v,w|\alpha,\beta,\pi,\rho,\tau)\ p(\alpha)\ p(\beta)\ p(\pi)\ p(\rho)\ p(\tau)\ d\alpha\ d\beta\ d\pi\ d\rho\ d\tau \\
= & \int p(x,y,|z,v,w,\alpha,\beta,\pi,\rho,\tau)\ p(z,v,w|\alpha,\beta,\pi,\rho,\tau)\ p(\alpha)\ p(\tau)\ p(\pi)\ p(\rho)\ p(\tau)\ d\alpha\ d\beta\ d\pi\ d\rho\ d\tau \\
= & \int p(x|z,v,w,\alpha,\beta,\pi,\rho,\tau)\ p(y|z,v,w,\alpha,\beta,\pi,\rho,\tau)\ p(z|\alpha,\beta,\pi,\rho,\tau) \\
& p(v,w|\alpha,\beta,\pi,\rho,\tau)\ p(\alpha)\ p(\beta)\ p(\pi)\ p(\rho)\ p(\tau)\ d\alpha\ d\beta\ d\pi\ d\rho\ d\tau \\
= & \int p(x|z,v,\alpha,\rho\ )\ p(y|z,w,\beta,\tau)\ p(z|\pi)\ p(v|\rho)\ p(w|\tau)\ p(\alpha)\ p(\beta)\ p(\pi)\ p(\rho)\ p(\tau)\ d\alpha\ d\beta\ d\pi\ d\rho\ d\tau \\
= & \int p(x|z,v,\alpha,\rho)\ d\alpha \int p(y|z,w,\beta,\tau)\ d\beta \int p(z|\pi)\ p(\pi)\ d\pi \int p(v|\rho)\ p(\rho)\ d\rho \int p(w|\rho)\ p(\tau)\ d\tau \\
= & \ p(x|z,v)\ p(y|z,w)\ p(z)\ p(v)\ p(w).
\end{aligned}
$$

D'où,

$$ICL(G,H,L) = \log p(x|z,v) + \log p(y|z,w) + \log p(z) + \log p(v) + \log p(w),$$

Il ne reste plus qu'à expliciter chaque terme pour obtenir le résultat suivant (voir section 7.4.1).

Nous supposons de manière analogue que les paramètres sont munis des lois a priori définies dans l'étape d'estimation. Nous pouvons alors calculer de manière exacte le critère ICL pour le modèle $MLBM$:

$$
\begin{aligned}
& ICL(G,H,L) \\
= & \ \log\Gamma(aG) + \log\Gamma(aH) + \log\Gamma(aL) - (G+H+L)\log\Gamma(a) \\
+ & \ G(H+L)(\log\Gamma(2b) - 2\log\Gamma(b)) - \log\Gamma(n+aG) - \log\Gamma(J+aH) - \log\Gamma(K+aL) \\
+ & \ \sum_{g=1}^{G}\log\Gamma(z_{\bullet g}+a) + \sum_{h=1}^{H}\log\Gamma(v_{\bullet h}+a) + \sum_{\ell=1}^{L}\log\Gamma(w_{\bullet \ell}+a) \\
+ & \ \sum_{g,h}\Big(\log\Gamma(n_g^h+b) + \log\Gamma(z_{\bullet g}v_{\bullet h} - n_g^h + b) - \log\Gamma(z_{\bullet g}v_{\bullet h}+2b)\Big) \\
+ & \ \sum_{g,\ell}\Big(\log\Gamma(n_g^\ell+b) + \log\Gamma(z_{\bullet g}w_{\bullet \ell} - n_g^\ell + b) - \log\Gamma(z_{\bullet g}w_{\bullet \ell}+2b)\Big),
\end{aligned}
\tag{7.1}
$$

où $n_g^h = \sum_{i,j} x_{ij}z_{ig}v_{jh}$ et $n_g^\ell = \sum_{i,j} x_{ij}z_{ig}w_{k\ell}$.

Dans la prolongation des résultats obtenus pour le LBM et pour le $MLBM$ dans la phase d'estimation, nous choisissons les mêmes hyperparamètres, à savoir $(a,b) = (4,1)$.

7.1.2 Critère Bayesian Information Criterion (BIC)

De manière analogue, nous pouvons proposer une expression du critère BIC en effectuant dans un premier temps, un développement asymptotique du critère ICL en n, J et K :

$$
\begin{aligned}
ICL^{Asympt}(G,H,L) \approx & \max_\theta \log p(x,y,\widehat{z},\widehat{v},\widehat{w};\theta) - \frac{G-1}{2}\log n - \frac{H-1}{2}\log J \\
& - \frac{L-1}{2}\log K - \frac{GH}{2}\log nJ - \frac{GL}{2}\log nK.
\end{aligned}
\tag{7.2}
$$

De même, le paramètre maximisant la log-vraisemblance complète intégrée est remplacé par l'estimateur MLE $\widehat{\theta}^{MLE}$ qui lui, est le paramètre maximisant la log-vraisemblance observée. Mais

ce dernier n'étant pas disponible, nous proposons de le remplacer par l'estimateur variationnel du maximum de l'énergie libre, approximation qui semble pertinente grâce au résultat de Brault et al. (2016) :

$$\begin{aligned}\max_{\theta} \log p(x, y, \widehat{z}, \widehat{v}, \widehat{w}; \theta) &\approx \log p(x, y, \widehat{z}, \widehat{v}, \widehat{w}; \widehat{\theta}) \\ &\approx \log p(\widehat{z}, \widehat{v}, \widehat{w} | x, y; \widehat{\theta}) + \log p(x, y; \widehat{\theta}) \\ &\approx \log p(\widehat{z}, \widehat{v}, \widehat{w} | x, y; \widehat{\theta}^{VB}) + \mathcal{F}(\widehat{\theta}^{VB}).\end{aligned}$$

Enfin, en utilisant la décomposition classique suivante :

$$ICL(G, H, L) = \log p(\widehat{z}, \widehat{v}, \widehat{w} | x, y; \widehat{\theta}) + BIC(G, H, L),$$

et en supposant la *conjecture* sur l'entropie énoncée dans la section 4.1.2, nous pouvons là aussi proposer l'heuristique suivante du critère BIC pour le modèle $MLBM$:

$$BIC(G, H, L) = \mathcal{F}(\widehat{\theta}^{VB}) - \frac{G(H+L)+G-1}{2} \log n - \frac{GH+H-1}{2} \log J - \frac{GL+L-1}{2} \log K \quad (7.3)$$

Remarque. Nous pouvons noter que $BIC(G, H, L) \neq BIC(G, L) + BIC(G, H)$ car le fait de considérer une partition en ligne commune à x et y donne une pénalité en $\log n$ plus petite.

7.2 Extension de la procédure Bi-KM1 au modèle $MLBM$

La procédure *Bi-KM1* présentée dans la section 4.2.2 peut être facilement étendue pour parcourir le nombre (G, H, L) de classes en ligne et en colonne de manière tri-directionnelle. De même, si nous nous plaçons sur une grille de taille $G_{\max} \times H_{\max} \times L_{\max}$, au lieu de parcourir $G_{\max} \times H_{\max} \times L_{\max}$ triplets, ce qui est très coûteux, nous parcourons $G_{\max} + H_{\max} + L_{\max}$ triplets de nombres. La figure 4.2 propose un exemple de représentation schématique de l'extension.

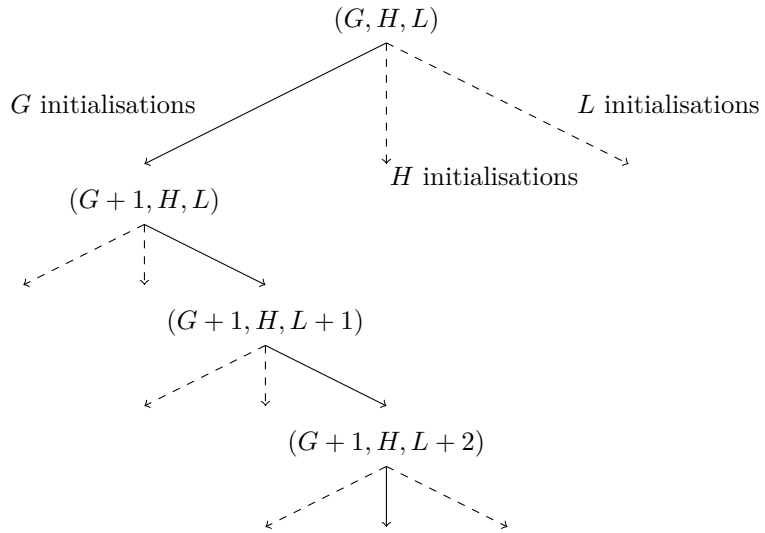


Figure 7.1 – Représentation schématique de l'algorithme *Bi-KM1* étendu.

7.3 Expérimentations numériques

Nous reprenons les matrices simulées au cours du plan d'expérience de la section 6.4.2. Cette expérience a été effectuée sur 50 matrices de chaque type.

Les résultats sont représentés dans les tableaux 7.1 et 7.2. Dans la plupart des cas, le critère ICL a sélectionné 3 classes en colonne pour la seconde matrice sauf dans le cas des matrices $+++$, où il y a un couple qui a été sélectionné *une fois* et qui est non présent sur les tableaux ci-dessus à savoir le couple $(3, 3, 2)$.

Ainsi, ICL semble avoir le même comportement dans le cas du LBM classique. Lorsque les données sont très mélangées (cas difficile $+++$), il a tendance à sous-estimer le nombre de classes.

$(1000, 100, 100)$

	$+, \ell = 3$			$++, \ell = 3$				$+++, \ell = 3$			
$g \backslash h$	4	5	6	2	3	4	5	2	3	4	5
3	0	0	0	0	0	0	0	0	28	0	0
4	27	0	0	0	0	49	0	0	1	20	0
5	**23**	0	0	0	0	**1**	0	0	0	**0**	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0

Table 7.1 – Répartition du nombre de classes en ligne et en colonne renvoyé par la procédure *Bi-KM1* étendu pour le critère ICL sur 50 matrices simulées pour chacune des difficultés ($+$ à $+++$) dans le cas de proportions déséquilibrées. Le rectangle représente le nombre de classes en ligne et en colonne qui a servi à générer les données.

$(1000, 100, 100)$

	$+, \ell = 3$			$++, \ell = 3$				$+++, \ell = 3$			
$g \backslash h$	4	5	6	2	3	4	5	2	3	4	5
3	0	0	0	0	0	0	0	0	4	0	0
4	2	0	0	0	0	50	0	0	0	45	0
5	**48**	0	0	0	0	**0**	0	0	0	**0**	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0

Table 7.2 – Répartition du nombre de classes en ligne et en colonne renvoyé par la procédure *Bi-KM1* étendu pour le critère ICL sur 50 matrices simulées pour chacune des difficultés ($+$ à $+++$) dans le cas de proportions équilibrées. Le rectangle représente le nombre de classes en ligne et en colonne qui a servi à générer les données.

Les résultats étant similaires, nous les présentons pour le critère BIC seulement dans le cas de proportions déséquilibrées en figure 7.3.

$+,\ \ell = 3$

$(1000, 100, 100)$	h 4	5	6
3	0	0	0
4	29	0	0
5	**21**	0	0
6	0	0	0
7	0	0	0

$++,\ \ell = 3$

g \ h	2	3	4	5
3	0	0	0	0
4	0	0	49	0
5	0	0	**1**	0
6	0	0	0	0
7	0	0	0	0

$+++,\ \ell = 3$

g \ h	2	3	4	5
3	0	31	0	0
4	0	0	19	0
5	0	0	**0**	0
6	0	0	0	0
7	0	0	0	0

Table 7.3 – Répartition du nombre de classes en ligne et en colonne renvoyé par la procédure *Bi-KM1* étendu pour le critère BIC sur 50 matrices simulées pour chacune des difficultés ($+$ à $+++$) dans le cas de proportions déséquilibrées. Le rectangle représente le nombre de classes en ligne et en colonne qui a servi à générer les données.

7.4 Annexes

7.4.1 Détails de la preuve de la formule (7.1)

Nous supposons donc que les distributions a priori de π, ρ et de τ sont des distributions de Dirichlet $\mathcal{D}(a, .., a)$.

De manière analogue, les calculs de $\log p(z)$, $\log p(v)$ et $\log p(w)$ étant similaires à ceux présentés dans la section 6.2.1.b, nous détaillons ici seulement les calculs de $p(x|z, v)$ et $p(y|z, w)$. Nous supposons que la loi a priori de α est une loi de Dirichlet $\mathcal{B}e(b, b)$. Comme les variables α et (z, v) sont indépendantes, nous avons

$$\begin{aligned} p(\alpha|x,z,v) &= \frac{p(x|z,v,\alpha)p(\alpha)}{\int p(x|z,v,\alpha)p(\alpha)d\alpha} \\ &= \prod_{g,h} \frac{(\alpha_{gh})^{n_g^h}(1-\alpha_{gh})^{z_{\cdot g}v_{\cdot h}-n_g^h}p(\alpha_{gh})}{\int (\alpha_{gh})^{n_g^h}(1-\alpha_{gh})^{z_{\cdot g}v_{\cdot h}-n_g^h}p(\alpha_{gh})d\alpha_{gh}} \\ &= \prod_{g,h} \frac{\binom{z_{+g}v_{+l}}{n_g^h}(\alpha_{gh})^{n_g^h}(1-\alpha_{gh})^{z_{\cdot g}v_{\cdot h}-n_g^h}p(\alpha_{gh})}{\int \binom{z_{\cdot g}v_{\cdot h}}{n_g^h}(\alpha_{gh})^{n_g^h}(1-\alpha_{gh})^{z_{\cdot g}v_{\cdot h}-n_g^h}p(\alpha_{gh})d\alpha_{gh}}, \end{aligned}$$

Enfin nous remarquons que la loi de α connaissant x, z, v est un produit de lois Beta $\mathcal{B}e(n_g^h + b, z_{\cdot g}v_{\cdot h} - n_g^h + b)$. Nous obtenons finalement que :

$$p(x|z,v) = \prod_{g,h} \frac{\Gamma(2b)}{\Gamma(b)^2} \frac{\Gamma(n_g^h + b)\Gamma(z_{\cdot g}v_{\cdot h} - n_g^h + b)}{\Gamma(z_{\cdot g}v_{\cdot h} + 2b)},$$

et en passant au logarithme, nous avons

$$\log p(x|z,v) = GH(\log\Gamma(2b) - 2\log\Gamma(b)) + \sum_{g,h}\Big(\log\Gamma(n_g^h + b) + \log\Gamma(z_{\cdot g}v_{\cdot h} - n_g^h + b) - \log\Gamma(z_{\cdot g}v_{\cdot h} + 2b)\Big)$$

De même, nous avons :

$$p(y|z,w) = \prod_{g,\ell} \frac{\Gamma(2b)}{\Gamma(b)^2} \frac{\Gamma(n_g^\ell + b)\Gamma(z_{\cdot g}w_{\cdot \ell} - n_g^\ell(y) + b)}{\Gamma(z_{\cdot g}w_{\cdot \ell} + 2b)},$$

et en passant au logarithme, nous obtenons également,

$$\log p(y|z,w) \quad = \quad GL(\log\Gamma(2b) - 2\log\Gamma(b)) + \sum_{g,\ell}\left(\log\Gamma(n_g^\ell + b) + \log\Gamma(z_{\bullet g}w_{\bullet \ell} - n_g^\ell + b) - \log\Gamma(z_{\bullet g}w_{\bullet \ell} + 2b)\right)$$

Le critère découle alors de ces trois formules. Nous pouvons remarquer d'un point de vue purement formel que

$$\begin{aligned} ICL_{MLBM}(x,y) &= ICL_{LBM}(x) + ICL_{LBM}(y) \\ &- \log\Gamma(aG) + G\log\Gamma(a) + \log\Gamma(n+aG) - \sum_{g=1}^{g}\log\Gamma(z_{+g}+a). \end{aligned}$$

7.4.2 Détails de la preuve de la formule (7.2)

Comme précédemment énoncé, le critère ICL s'écrit :

$$ICL(G,H,L) = \log p(x,y,\widehat{z},\widehat{v},\widehat{w}) = \log p(x|\widehat{z},\widehat{v}) + \log p(y|\widehat{z},\widehat{w}) + \log p(\widehat{z}) + \log p(\widehat{v}) + \log p(\widehat{w})$$

Une approximation asymptotique de chaque terme est réalisée dans l'égalité précédente en utilisant toujours l'approximation suivante :

$$\log\Gamma(z) \underset{z\to+\infty}{\sim} (z-\frac{1}{2})\log z - z - \frac{1}{2}\log 2\pi.$$

Les calculs étant similaires à la section précédente, nous donnons directement l'expression trouvée pour $\log p(\widehat{z})$, $\log p(\widehat{v})$ et $\log p(\widehat{w})$:

$$\log p(\widehat{z}) \sim \max_{\pi}\log p(\widehat{z};\pi) - \frac{G-1}{2}\log n.$$

De même, on a

$$\log p(\widehat{v}) \sim \max_{\rho}\log p(\widehat{v};\rho) - \frac{H-1}{2}\log J,$$

et enfin,

$$\log p(\widehat{w}) \sim \max_{\tau}\log p(\widehat{w};\tau) - \frac{L-1}{2}\log K.$$

Maintenant regardons,

$$\log p(x|\widehat{z},\widehat{v}) \quad = \quad GH(\log\Gamma(2b) - 2\log\Gamma(b)) + \sum_{g,h}\left(\underbrace{\log\Gamma(\widehat{n_g^h}+b)}_{A} + \underbrace{\log\Gamma(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}} - \widehat{n_g^h} + b)}_{B} - \underbrace{\log\Gamma(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}} + 2b)}_{C}\right)$$

Alors

$$\begin{aligned} A &\sim \sum_{gh}\left((\widehat{n_g^h}+b-\frac{1}{2})\log(\widehat{n_g^h}+b)-(\widehat{n_g^h}+b)\right) \\ &\sim \sum_{gh}\left(\widehat{n_g^h}\log\widehat{n_g^h}+\cancel{b}+(b-\frac{1}{2})\log\widehat{n_g^h}-\widehat{n_g^h}-\cancel{b}\right) \\ &\sim \sum_{gh}[\widehat{n_g^h}\log\widehat{n_g^h}]+GH(b-\frac{1}{2})\log nJ-GHnJp, \end{aligned}$$

où $\widehat{n_g^h}=\sum_{i,j}x_{ij}\widehat{z_{ig}}\widehat{v_{jh}}=nJp$, avec p la proportion de cellules noires.

Et nous avons également,

$$\begin{aligned} B &\sim \sum_{g,h}\left((\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h}+b-\frac{1}{2})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h}+b)-(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h}+b)\right) \\ B &\sim \sum_{g,h}\left((\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})+\cancel{b}+(b-\frac{1}{2})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})-\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}+\widehat{n_g^h}-\cancel{b}\right) \\ B &\sim \sum_{g,h}[(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})-\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}+\widehat{n_g^h}]+(b-\frac{1}{2})\sum_{g,h}\log[nJ(1-p)] \\ B &\sim \sum_{g,h}[(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})-nJ(1-p)]+GH(b-\frac{1}{2})\log nJ, \end{aligned}$$

car $\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h}=\sum_{i,j}x_{ij}\widehat{z_{ig}}\widehat{v_{jh}}(1-x_{ij})=nJ(1-p)$.

Enfin,

$$\begin{aligned} C &\sim \sum_{g,h}\left((\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}+2b-\frac{1}{2})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}+2b)-(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}+2b)\right) \\ C &\sim \sum_{g,h}\left(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}})+\frac{\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}}{\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}}2b+(2b-\frac{1}{2})\log\widehat{z_{\bullet g}}\widehat{v_{\bullet h}})-\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\cancel{2b}\right) \\ C &\sim \sum_{g,h}[\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}})]+GH(2b-\frac{1}{2})\log nJ-GHnJ \end{aligned}$$

Nous pouvons conclure alors que

$$\begin{aligned} A+B-C &\sim \sum_{g,h}[\widehat{n_g^h}\log\widehat{n_g^h}]+GH(b-\frac{1}{2})\log nJ-\cancel{GHnJp} \\ &+ \sum_{g,h}[(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}-\widehat{n_g^h})]-\cancel{GHnJ(1-p)}+GH(b-\frac{1}{2})\log nJ \\ &- \sum_{g,h}[\widehat{z_{\bullet g}}\widehat{v_{\bullet h}}\log(\widehat{z_{\bullet g}}\widehat{v_{\bullet h}})]-GH(2b-\frac{1}{2})\log nJ+\cancel{GHnJ} \\ &\sim GH\log nJ[\cancel{b}-\frac{1}{2}+\cancel{b}-\frac{\cancel{1}}{\cancel{2}}-\cancel{2b}+\frac{\cancel{1}}{\cancel{2}}] \\ &\sim \max_{\alpha}\log p(x|\widehat{z},\widehat{v};\alpha)-\frac{GH}{2}\log nJ, \end{aligned}$$

car

$$\max_{\alpha} \log p(x|\widehat{z},\widehat{v};\alpha) \quad = \quad \sum_{g,h}[\widehat{n_g^h}\log\widehat{n_g^h}] + \sum_{g,h}[(\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h})\log(\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h})] - \sum_{g,h}[\widehat{z_{.g}}\widehat{v_{.h}}\log(\widehat{z_{.g}}\widehat{v_{.h}})].$$

En effet, $\max_{\alpha} p(x|\widehat{z},\widehat{v};\alpha) = \prod_{i,j,g,h}[\widehat{\alpha_{gh}}^{x_{ij}}(1-\widehat{\alpha_{gh}})^{1-x_{ij}}]^{\widehat{z_{ig}}\widehat{v_{jh}}}$, avec $\widehat{\alpha_{gh}} = \frac{\widehat{n_g^h}}{\widehat{z_{+g}}\widehat{v_{+h}}}$.

D'où

$$\begin{aligned}
\max_{\alpha} \log p(x|\widehat{z},\widehat{v};\alpha) &= \sum_{g,h}\widehat{n_g^h}\log\frac{\widehat{n_g^h}}{\widehat{z_{.g}}\widehat{v_{.h}}} + (\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h})\log\frac{\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h}}{\widehat{z_{.g}}\widehat{v_{.h}}} \\
&= \sum_{g,h}\widehat{n_g^h}\log\widehat{n_g^h} + \sum_{g,h}(\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h})\log(\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h}) \\
&- \cancel{\sum_{g,h}\widehat{n_g^h}\log\widehat{z_{.g}}\widehat{v_{.h}}} - \cancel{\sum_{g,h}(\widehat{z_{.g}}\widehat{v_{.h}} - \widehat{n_g^h})\log\widehat{z_{.g}}\widehat{v_{.h}}}.
\end{aligned}$$

Ainsi, ce terme est équivalent à

$$\log p(x|\widehat{z},\widehat{v}) \sim \max_{\alpha}\log p(x|\widehat{z},\widehat{v};\alpha) - \frac{GH}{2}\log nJ.$$

De manière analogue, on a

$$\log p(y|\widehat{z},\widehat{w}) \sim \max_{\beta}\log p(y|\widehat{z},\widehat{w};\beta) - \frac{GL}{2}\log nK.$$

Pour conclure et retrouver l'approximation annoncée, nous utilisons le lemme ci-dessous dont la démonstration est laissée au lecteur :

Lemme 7.4.2.1.

$$\begin{aligned}
\max_{\theta}\log p(x,y,\widehat{z},\widehat{v},\widehat{w};\theta) &= \max_{\pi}\log p(\widehat{z};\pi) + \max_{\rho}\log p(\widehat{v};\rho) + \max_{\tau}\log p(\widehat{w};\tau) \\
&+ \max_{\alpha}\log p(x|\widehat{z},\widehat{v};\alpha) + \max_{\beta}\log p(y|\widehat{z},\widehat{w};\beta).
\end{aligned}$$

Part III

Conclusion et Perspectives

8

Méthodologie proposée pour l'analyse des données individuelles de pharmacovigilance

Ce chapitre résume l'ensemble du travail réalisé dans cette thèse en proposant une méthode pour analyser les données individuelles de pharmacovigilance. Nous commençons par présenter la méthodologie : dans un premier temps, nous étudions le tableau de contingence "médicaments-effets indésirables" à l'aide du modèle des blocs latents exposé dans la partie I de cette thèse. Cette analyse fournit une liste pertinente de groupes de médicaments et de groupes d'effets indésirables. À partir de cette liste, nous effectuons une analyse des données individuelles à l'aide du modèle des blocs latents multiple développé en partie II qui permet de prendre en compte l'interaction entre plusieurs effets indésirables et de pallier le phénomène de coprescription. Dans un second temps, nous proposons de tester cette méthodologie sur des données simulées à partir d'un modèle de simulation numérique que nous avons développé et que nous détaillerons. Puis, nous illustrons cette méthodologie sur un jeu de données réelles : la base française de pharmacovigilance entre 2002 et 2010, administrée par l'AFFSAPS, non accessible librement mais rendue disponible par l'équipe B2PHI (UMR 1181, INSERM, Villejuif) que nous remercions chaleureusement.

8.1 Méthodologie proposée pour analyser les données individuelles en pharmacovigilance

8.1.1 Problématique

Les données individuelles de pharmacovigilance ne sont pas directement analysées actuellement. À la place, les praticiens utilisent un tableau de contingence résumant ces données individuelles puis cherchent à détecter des associations "médicaments-effets indésirables" à partir de ce tableau uniquement (*Proportional Reporting Ratio* (PRR, Evans et al. (2001)), *Reporting Odds Ratio* (ROR, van Puijenbroek et al. (2002)), *Bayesian Confidence Propagation Neural Network* ($BCPNN$, Bate et al. (1998); Norén et al. (2006)), et *(Multi-item) Gamma Poisson Shrinker* ($(M)GPS$, DuMouchel (1999); DuMouchel and Pregibon (2001)). Or, l'analyse directe des données individuelles permettrait de mettre à jour des "effets de coprescription " entre médicaments, ce qui est impossible à détecter à l'aide du tableau de contingence. Le phénomène de coprescription est le suivant : un médicament peut être souvent coprescrit avec un médicament impliqué dans un signal, et l'effet indésirable peut être imputé à tort à ce médicament (Caster et al. (2010)). C'est pour cela que les données individuelles s'avèrent très utiles pour contourner ce problème. Mais au vu de leur taille (production de matrices de taille environ 200 000 $\times$ 2000 et de taille 200 000 $\times$ 4000) qui rend les traitements informatiques de celles-ci difficiles, nous proposons une méthodologie en deux étapes.

8.1.2 Résumé de la méthode proposée

Nous proposons donc la méthode suivante :

- classification croisée du tableau de contingence via le modèle LBM Poisson normalisé,
- classification croisée des données individuelles via le modèle $MLBM$ en utilisant l'information a priori issue de la classification croisée du tableau de contingence.

8.2 Un modèle de simulation numérique

Dans cette section, nous illustrons la méthode proposée sur des données simulées. Nous expliquons d'abord comment nous simulons ces données afin de pouvoir comparer ensuite, les méthodes existantes entre elles.

8.2.1 Plan de simulation

Pour réaliser cette simulation, nous procédons en deux temps :

- simulation de la matrice x de prise des médicaments, en supposant que les individus sont tous indépendants les uns des autres.
- Puis, à partir de cette matrice x, simulation de la matrice y de survenue des effets indésirables en incorporant des signaux.

8.2.1.a Matrice x de prise de médicaments

D'abord, nous considérons la matrice x de taille $n \times J$ composée uniquement de 0 et nous supposons que les prises de médicaments entre deux individus i et i' sont indépendantes ; ceci nous permet alors de faire une simulation ligne par ligne. Étant donné un individu i allant de 1 à n, nous devons

- simuler le nombre N_i de médicaments qu'a pris l'individu i,
- puis choisir les N_i médicaments parmi les J médicaments présents.

Ainsi, pour chaque individu i nous procédons en deux étapes :

1. Pour la simulation de la variable N_i, nous choisissons de prendre la distribution empirique associée aux données réelles (voir figure 8.1) c'est-à-dire que pour tout m entier naturel non nul, nous calculons

$$w_m = \frac{\text{Nombre d'individus dans la base réelle ayant pris } m \text{ médicaments}}{\text{Nombre total d'individus dans la base réelle}}$$

et ainsi, nous simulons N_i tel que pour tout $m \in \mathbb{N}^\star$, nous ayons :

$$\mathbb{P}(N_i = m) = w_m.$$

2. Une fois le nombre de médicaments fixé pour l'individu i, nous choisissons les médicaments pris ; autrement dit, nous cherchons les indices j compris entre 1 et J tels que x_{ij} vaille 1. Pour cela, nous nous sommes basés sur la forme d'une loi géométrique de paramètre p ; nous calculons pour tout $j \in \{1, \dots, J\}$

$$\omega_j = \frac{(1-p)^j p}{\sum_{j'=1}^{J} (1-p)^{j'} p} = \frac{(1-p)^j p}{1-(1-p)^J},$$

puis nous faisons un tirage sans remise de N_i médicaments parmi J, où chaque médicament j a un poids ω_j d'être choisi ; dans ce cas, nous donnons à la cellule x_{ij} la valeur 1. Ainsi, nous favorisons le médicament 1 à être le plus pris, puis le médicament 2, etc...

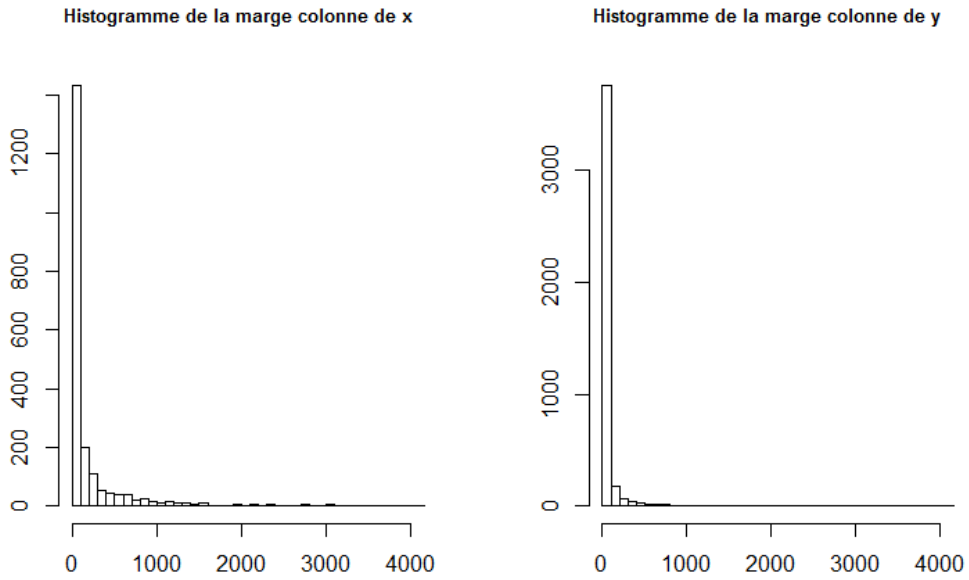


Figure 8.1 – Histogrammes des marges colonnes des données réelles x (à gauche) et de y (à droite).

Une fois les simulations faites pour chaque individu, nous obtenons une matrice x où chaque ligne i possède exactement N_i cellules valant 1 et que des 0. Il est possible qu'une fois cette étape terminée, il existe un médicament j qui n'ait été pris par aucun individu (dans ce cas, il ne devrait pas apparaître dans la base) ; nous tirons alors aléatoirement un individu i dans la base de données

et nous supposons qu'il prend ce médicament ($x_{ij} = 1$).

Enfin, cette simulation induit que le premier médicament est celui qui a été pris le plus, puis vient le deuxième et ainsi de suite. Pour pallier ce phénomène, nous faisons une permutation aléatoire des colonnes.

Remarque. La valeur p a été calibrée manuellement afin que la courbe marginale représentant la somme des individus ayant pris chaque médicament ressemble le plus possible aux données réelles (voir figure 8.1). Ce paramètre p dépend du nombre n d'individus et du nombre J de médicaments présents dans la base (ici par exemple p≃0.08 pour $n = 20000$ et $J = 200$).

8.2.1.b Matrice y de survenue des effets

Pour la simulation des effets indésirables, nous avons fait plusieurs hypothèses :

- nous supposons que pour chaque individu, il y a au moins un médicament et un effet indésirable recensé, comme c'est le cas pour la base de données réelles.
- Il est possible qu'un individu ait pris deux médicaments et que chacun de ces médicaments ait entrainé le même effet indésirable. Dans ce cas, le fait que l'effet a été causé par plusieurs médicaments est ignoré.
- Si un médicament est impliqué dans un signal, le fait de prendre ce médicament augmente fortement les probabilités d'avoir l'effet.
- Il est possible de prendre un médicament non impliqué dans un signal et d'avoir tout de même un effet indésirable (provenant d'une cause extérieure).

Pour schématiser les deux dernières hypothèses, nous introduisons la *matrice de génération des signaux* S de taille $J \times K$ qui comporte exactement n_S valeurs valant 1. Ainsi, nous avons pour $j \in \{1, \dots, J\}$ et $k \in \{1, \dots, K\}$ que

$$S_{jk} = \begin{cases} 1 & \text{si le couple (médicament } j\text{, effet } k\text{) est un signal,} \\ 0 & \text{sinon.} \end{cases}$$

Notons que cette matrice S résume l'information que recherchent toutes les méthodes de pharmacovigilance. Pour la suite, nous la simulons en supposant qu'elle ne contient que des 0 sauf pour n_S cellules tirées au hasard ; le tirage se fait de façon équiprobable entre toutes les cellules de la matrice.

Pour introduire la partie aléatoire des deux dernières hypothèses, nous créons la matrice $S^q = q \times S + \varepsilon(1 - S)$, ou autrement dit, nous avons pour $j \in \{1, \dots, J\}$ et $k \in \{1, \dots, K\}$ que

$$S^q_{jk} = \begin{cases} q & \text{si le couple (médicament } j\text{, effet } k\text{) est un signal,} \\ \varepsilon & \text{sinon.} \end{cases}$$

Par la suite, cette matrice S^q va nous permettre de simuler l'apparition de l'effet k sachant qu'un individu ait pris un médicament j à l'aide d'une loi de Bernoulli de paramètre S^q_{jk}. Notons donc deux faits :

- Plus q sera proche de 1 et plus les signaux seront présents dans la base.
- Plus ε sera grand et plus il y aura des effets indésirables apparaissant sans raisons valables.

Maintenant que nous avons généré la matrice S^q, il nous faut simuler la matrice des effets indésirables. Pour cela, nous considérons également une matrice y nulle mais cette fois de taille $n \times K$. Comme précédemment, nous supposons que la survenue d'effets indésirables pour l'individu i est indépendante de ceux obtenus par l'individu i' ; ce qui nous permet à nouveau de faire la simulation ligne par ligne. Ainsi, pour chaque individu i allant de 1 à n, nous effectuons la procédure suivante :

1. Nous commençons par regarder les médicaments $j_1, j_2, \ldots, j_{N_i}$ qu'il a pris puis pour chaque effet k allant de 1 à K :
   (a) Nous calculons N_i variables de Bernoulli $Z_{j_1}, \ldots, Z_{j_{N_i}}$ telles que pour tout $j \in \{j_1, j_2, \ldots, j_{N_i}\}$
   $$Z_j \sim \mathcal{B}\left(S^q_{jk}\right).$$
   (b) Ensuite, nous affectons à la variable y_{ik} la valeur maximum des variables obtenues :
   $$y_{ik} = \max\left(Z_{j_1}, \ldots, Z_{j_{N_i}}\right).$$
   Ainsi, si plusieurs médicaments entraînent l'effet indésirable k, plusieurs variables Z seront égales à 1 mais nous n'aurons pas cette information dans la matrice finale.
2. Si à l'issue de l'étape précédente, l'individu i ne possède aucun effet indésirable (ce qui peut arriver si les médicaments pris ne font partie d'aucun signal), nous avons deux possibilités :
   - Reprendre l'étape 1 jusqu'à obtention d'au moins un effet indésirable ; ce qui peut être coûteux en temps, si l'individu n'a pris qu'un seul médicament non impliqué dans un signal et si ε est très petit.
   - Choisir un effet indésirable au hasard et supposer que l'individu i l'a obtenu ; ce qui peut créer un biais en faisant augmenter fortement le nombre d'individus n'ayant eu qu'un seul effet indésirable.

À la fin, nous obtenons une matrice y avec au moins une cellule y_{ik} valant 1 par ligne. Comme pour les médicaments, il est possible qu'un effet indésirable ne soit jamais notifié. Dans ce cas, nous choisissons à nouveau un individu au hasard et nous supposons qu'il a eu l'effet.

Remarque. Notons que pour la prise d'un seul médicament, nous testons à chaque fois tous les effets indésirables. Ainsi, si nous admettons que ce médicament n'est présent dans aucun signal, la probabilité qu'il n'entraîne aucun effet indésirable est de $(1-\varepsilon)^K$. *Par conséquent, à* ε *fixé, lorsque* K *augmente, le bruit augmente également (puisque la probabilité de ne pas avoir d'effets indésirables provenant de ce médicament tend vers 0). C'est pourquoi, nous sommes obligés de prendre des valeurs pour* ε *très petites dans les simulations afin de conserver une matrice relativement creuse. Nous constatons donc que la difficulté va dépendre à la fois de* q *et de* ε *mais aussi de* K.

8.2.2 Expérimentations numériques et évaluation via les courbes ROC

8.2.2.a Rappel des méthodes étudiées

Nous étudions les méthodes présentées dans le chapitre 2 et qui sont utilisées officiellement dans l'un des systèmes de pharmacovigilance européen ou américain. Nous rappelons qu'elles sont basées sur le tableau de contingence c et à partir de celui-ci, pour chaque couple (médicament/effet) présent dans la base, un tableau de contingence 2×2 est élaboré (voir tableau 8.1).

	Médicament j	Autres médicaments	Marges lignes
Effet indésirable k	c_{jk}=a	b	$c_{+k} = a + b$
Autres effets	c	d	$c + d$
Marges colonnes	$c_{j+} = a + c$	$b + d$	c_{tot}

Table 8.1 – Tableau de contingence 2×2 pour chaque couple (médicament/effet) présent dans la base.

Les méthodes alors considérées sont :

- **PRR (Proportionnal Reporting Ratio)** : $PRR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$.

- **ROR (Reporting Odds Ratio)** : $ROR = \frac{ad}{bc}$.
- **IC (Information Component**, Bate (98)):
  $IC_{jk} = \log_2(\frac{\frac{n_{jk}}{N}}{\frac{n_{+k}}{N} \times \frac{n_{j+}}{N}})$,
- **Méthode GPS (Gamma Poisson Shrinker**, Dumouchel (99))
  $GPS_{jk} = \log_2(\frac{\text{observé}}{\text{attendu}})$, où $\frac{\text{observé}}{\text{attendu}} \sim$ mélange de 2 lois Γ.

Dans tout ce qui suit, les expérimentations sur ces méthodes ont été réalisées en utilisant le package du logiciel R intitulé *PhViD: PharmacoVigilance Signal Detection* (Ahmed et al. (2010)).

8.2.2.b Expérimentations numériques avec q=0.8

Dans cette partie, nous simulons différentes matrices suivant le modèle précédent. Pour la taille, nous avons cherché à prendre des valeurs permettant de se rapprocher de la forme des données réelles. Pour cela, nous avons choisi $(n, J, K) = (20\,000, 200, 400)$.

Pour les autres paramètres, nous supposons qu'il y a $n_S = 300$ signaux à découvrir (soit une proportion de 0.375%). Notons alors que certains médicaments peuvent ne pas être impliqués dans un signal. Nous fixons q à 0.8 ce qui signifie que, pour chaque couple médicament/effet présent dans un signal, lorsque 5 individus prennent le médicament, en moyenne 4 d'entre eux ont l'effet associé.

Pour le bruit, nous avons pris ε dans $\{10^{-1}, 10^{-2}, \ldots, 10^{-6}\}$. En comparaison de la valeur q, nous pouvons penser que ces chiffres sont très petits mais lorsqu'un individu prend un médicament, la probabilité de ne pas avoir du tout d'effet indésirable non présent dans un signal est de $(1-\varepsilon)^K$. Dans le cas où $\varepsilon = 10^{-1}$, cela représente une probabilité d'environ 5×10^{-19} et ceci pour chaque médicament pris par chaque individu.

Courbes ROC

Les estimations renvoyées par les méthodes précédentes dépendent d'un seuil que nous pouvons faire varier. Nous avons donc décidé de comparer les méthodes à l'aide de courbes ROC (Receiver Operating Characteristic) ; voir les figures 8.2, 8.3 et 8.4. Nous remarquons que dans le cas où la puissance du signal est assez forte avec un niveau de bruit variable, les méthodes étudiées recouvrent très bien les signaux générés, et les méthodes ROR et PRR semblent les plus performantes dans cette modélisation. De plus, les seuils préconisés par défaut semblent bien être calibrés si l'objectif est de minimiser le nombre de faux positifs. Remarquons en revanche que dans la réalité, les méthodes GPS et BCPNN sont considérées comme les plus performantes (voir section 8.3.2).

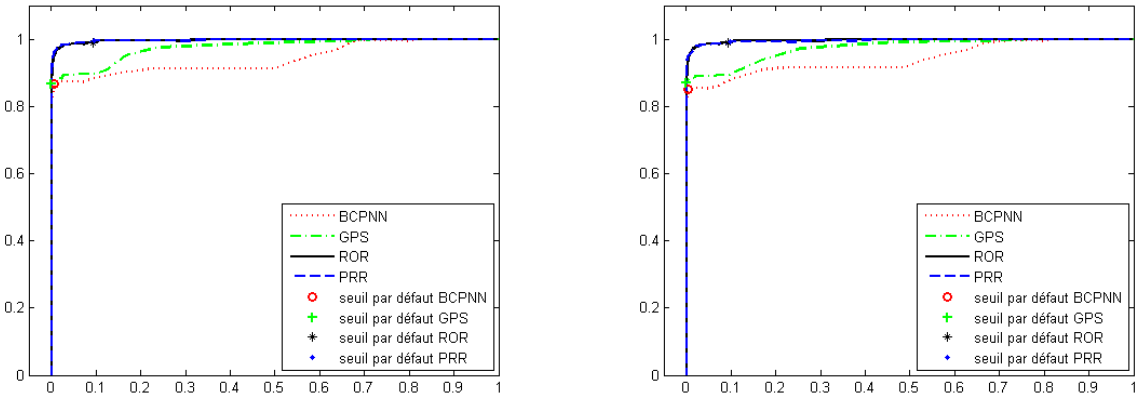


Figure 8.2 – Courbes ROC pour $\varepsilon = 10^{-6}$ à gauche et $\varepsilon = 10^{-5}$ à droite.

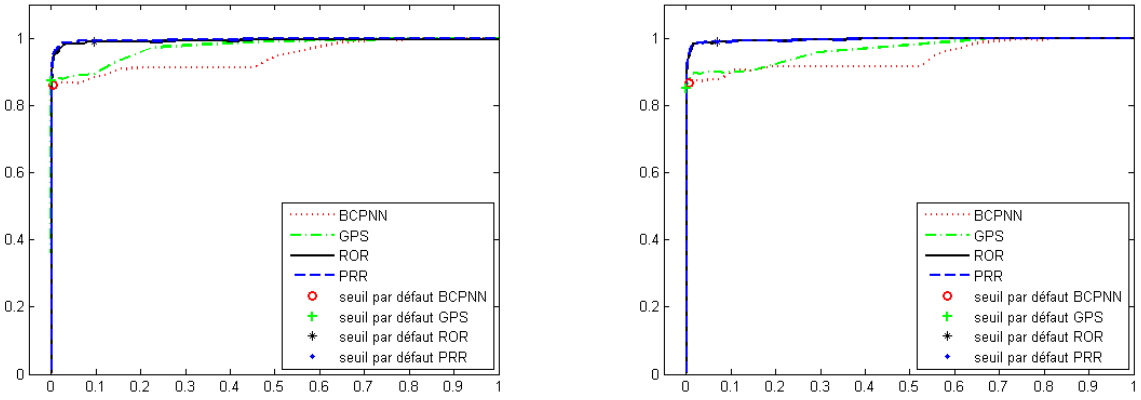


Figure 8.3 – Courbes ROC pour $\varepsilon = 10^{-4}$ (à gauche) et $\varepsilon = 10^{-3}$ (à droite).

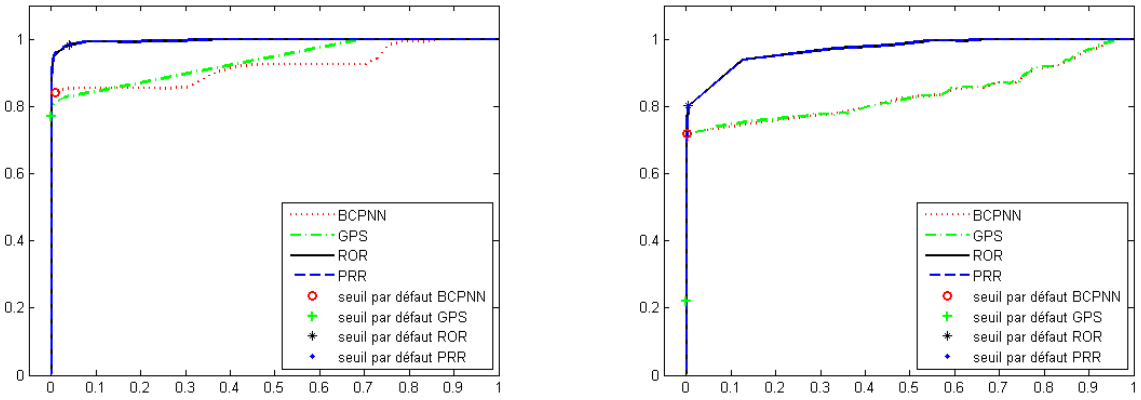


Figure 8.4 – Courbes ROC pour $\varepsilon = 10^{-2}$ (à gauche) et $\varepsilon = 10^{-1}$ (à droite).

Dans cette partie, nous montrons sur un exemple comment analyser les données à partir des modèles présentés en chapitres 3 et 6. Nous commençons par remarquer qu'une analyse faite uniquement sur les données individuelles serait assez coûteuse en temps pour deux raisons :

- il y a trois paramètres de classes à estimer (G, H, L) et donc à faire varier.
- Le nombre d'individus étant très important, les matrices x et y ont des tailles très grandes ; ce qui peut ralentir fortement l'application des algorithmes.

Pour contourner ces difficultés, nous avons opté pour une analyse en deux temps :

1. D'abord, nous utilisons le modèle des blocs latents sur le tableau de contingence $c = t(x)y$ du chapitre 3, dont la taille est beaucoup plus réduite, afin d'extraire un couple $(H_{\min}, L_{\min})$ de nombres minimaux de groupes pour les médicaments et pour les effets ainsi que les classes associées. En plus d'une réduction drastique du temps nécessaire à l'obtention de ces classes, nous pouvons espérer que l'algorithme commence à rassembler des signaux ensembles.
2. Ensuite, à l'aide des classes obtenues, nous utilisons le modèle $MLBM$ (chapitre 6) en initialisant le nombre de classes à $(2, H_{\min}, L_{\min})$, en prenant pour les classes de médicaments et d'effets indésirables celles obtenues précédemment et en utilisant un algorithme des k-means sur les lignes afin de former les deux premières classes d'individus. Pour l'algorithme des k-means, nous avions le choix de tenir compte des classes formées en colonne (ce qui donnerait plus de poids encore aux résultats du LBM pour le tableau de contingence) ou non. C'est ce deuxième choix qui a été conservé dans cet exemple.

   Pour l'analyse des données individuelles, nous utilisons la procédure *Bi-KM1* étendue. Nous nous attendons à ce que, dans un premier temps, ce soit principalement le nombre de classes en ligne qui évolue. De plus, comme nous ne figeons pas les classes en colonnes, il est tout à fait possible que celles-ci soient différentes au fur et à mesure que les classes des individus se forment.

Pour analyser cette procédure étape par étape, nous avons repris le plan de simulation précédent $((n, J, K, n_S) = (20\,000, 200, 400, 300))$ avec $\varepsilon = 10^{-6}$ et nous détaillons ici les résultats obtenus.

L'analyse sur le tableau de contingence a fourni 27 classes de médicaments et 29 classes d'effets indésirables. Sur la figure 8.5, nous avons représenté la matrice réorganisée ainsi que les emplacements des 300 signaux. Nous pouvons remarquer que, même si le modèle qui a servi a généré les données n'est pas celui du LBM, l'algorithme a tout de même eu tendance à créer des blocs regroupant un grand nombre de signaux.

Nous nous sommes ensuite interrogés sur ce qui caractérisait ces blocs et nous avons remarqué que les blocs ayant les intensités $\gamma_{h\ell}$ les plus élevées correspondaient à ceux qui comportaient beaucoup de signaux. Nous avons représenté sur le tableau 8.2 la liste des dix premiers blocs suivant cet ordre et nous remarquons que ceux-ci possèdent des proportions de signaux allant de 13% à 50%. Rappelons que, si les blocs avaient été formés au hasard, ces proportions se situeraient autour de 0,375%. Il y a donc entre 35 et 133 fois plus de signaux présents en moyenne dans ces blocs particuliers.

Avec les résultats obtenus, nous avons utilisé l'algorithme *Bi-KM1* sur les tableaux x et y des données individuelles. Le critère ICL a sélectionné le triplet $(50, 32, 40)$. Nous pouvons remarquer que les nombres de classes des médicaments et des effets indésirables ont donc augmenté. Si nous avions eu les vraies données réelles, nous aurions pu chercher à établir des profils des individus mis dans les mêmes classes (Est-ce des personnes âgées ? Des femmes enceintes ?).

Comme nous sommes dans le cadre de données simulées, nous ne pouvons qu'essayer de comprendre comment l'algorithme a séparé ces classes. En représentant la matrice (voir figure 8.6), nous pouvons constater que certains individus sont regroupés car ils ont tendance à prendre les mêmes médicaments et à avoir les mêmes effets indésirables. En particulier, nous avons envie de faire correspondre les blocs de forte intensité de la matrice des médicaments avec les blocs de forte intensité de la matrice des effets afin de vérifier s'il n'y a pas des signaux dans les couples possibles.

En effet, pour préciser cette analyse, nous avons regardé pour chaque groupe d'individus quelles étaient les classes de médicaments et d'effets dont les intensités α et β étaient les plus élevées. Dans le tableau 8.3, nous avons représenté les 20 premières classes d'individus avec les couples (α,β) d'intensité la plus forte ainsi que les classes de médicaments et d'effets associées, les effectifs de chaque classe et si parmi les médicaments et effets présents, il y avait des signaux générés présents.

Notons que pour ces valeurs fortes de α et β, nous avons peu de médicaments et d'effets dans les classes correspondantes, mais qu'ils sont presque systématiquement associés à des signaux ; nous voyons même que pour la classe d'individus $g = 43$, nous avons isolé un même médicament présent dans deux signaux.

De plus, la procédure a tendance à séparer les individus qui ont pris le médicament et qui ont eu les effets, de ceux qui ont pris le médicament et qui ne les ont pas eus. En séparant de la sorte ces individus, nous pouvons espérer détecter plus facilement les signaux.

Pour les données réelles, cela pourrait permettre par ailleurs, de donner des profils de personnes à risque pour les effets secondaires (enfants, personnes âgées, femmes enceintes...). Ce comportement pourrait aussi mettre en évidence les effets indésirables causés par la prise de deux médicaments joints puisque les individus qui n'en prendraient qu'un seul et qui n'auraient pas l'effet du coup seraient écartés dans une autre classe. Bien que la méthode soit perfectible puisqu'il faut encore trouver un moyen d'automatiser la détection, nous constatons que nous mettons en évidence un grand nombre de signaux.

Enfin, notons qu'une limitation de notre modèle aurait été l'obtention de groupes d'individus très petits qui n'aurait pas permis une analyse des résultats (comment interpréter le fait qu'un seul individu soit dans une classe ?). Toutefois, la valeur de $a = 4$ pour l'a priori sur les classes permet de limiter les petites classes puisque sur ce jeu de 20 000 individus, seules deux classes ont un peu moins d'une centaine de lignes.

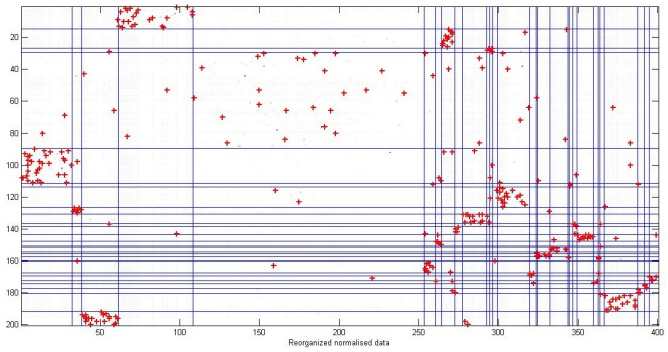


Figure 8.5 – Tableau de contingence réorganisé à l'aide des partitions estimées. Les croix rouges représentent les signaux.

γ	numéro classe en ligne	numéro classe en colonne	proportion de signaux par bloc en %
3.7285e-04	12	22	25
2.8941e-04	6	20	50
1.8891e-04	23	23	33
1.8674e-04	10	21	50
1.2535e-04	25	27	25
1.0138e-04	14	25	33
6,0865e-05	6	27	13
5,2487e-05	17	17	38
4,6917e-05	24	28	33
4,5367e-05	15	18	20

Table 8.2 – Les 10 blocs avec les intensités γ les plus fortes rangés par ordre décroissant.

n^o classe d'individus	plus grande valeur de α	plus grande valeur de β	classe de médicaments correspondante	classe d'effets correspondante	effectif des individus	nombre de médicaments	nombre d'effets	nombre de signaux	proportion de signaux (%)
42	1	1	31	24	245	1	1	1	100
43	1	0,88	18	31	123	1	2	2	100
23	1	0,87	2	24	108	1	1	1	100
37	1	0,87	18	31	133	1	2	2	100
35	1	0,87	30	19	136	1	1	1	100
46	1	0,83	13	37	177	1	1	1	100
29	1	0,82	18	31	244	1	2	2	100
1	1	0,82	14	17	90	1	3	3	100
44	1	0,82	14	17	180	1	3	3	100
41	1	0,82	18	31	208	1	2	2	100
49	1	0,79	22	29	108	1	2	2	100
21	0,99	0,83	22	29	222	1	2	2	100
9	0,99	0,89	30	11	224	1	1	1	100
10	0,99	0,83	22	29	319	1	2	2	100
34	0,99	0,79	13	36	312	1	4	4	100
28	0,99	0,84	18	31	210	1	2	2	100
5	0,99	0,85	22	29	453	1	2	2	100
24	0,99	0,81	22	29	288	1	2	2	100
45	0,99	1	31	24	352	1	1	1	100
38	0,97	0,80	22	29	167	1	2	2	100

Table 8.3 – Les 20 premiers groupes d'individus rangés par ordre décroissant des valeurs de α les plus fortes pour chaque groupe d'individus.

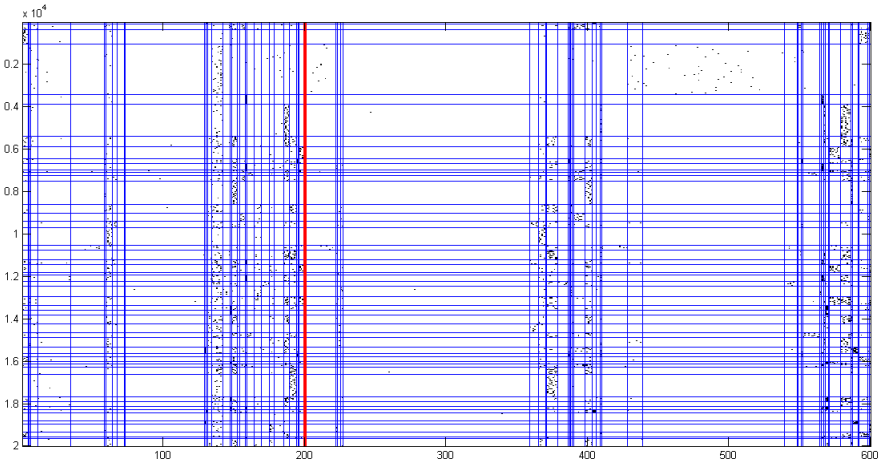


Figure 8.6 – Matrices x et y réorganisées à l'aide des partitions estimées. Celles-ci sont séparées par un trait rouge.

8.2.2.c Exemple de simulation de la coprescription

SENSIBILITÉ DES MÉTHODES USUELLES À LA PROBLÉMATIQUE DE LA COPRESCRIPTION

L'un des intérêts du modèle de simulation présenté en section 8.2 est de pouvoir l'étendre aisément à la problématique de la coprescription. Celle-ci réside dans le fait suivant : un médicament B non impliqué dans un signal peut être souvent coprescrit avec un médicament A impliqué dans un signal, et l'effet indésirable peut être imputé à tort à ce médicament B par les méthodes usuelles utilisant l'information du tableau de contingence. En effet, sur le tableau de contingence, l'information résumée montre que l'effet a été notifié un grand nombre de fois avec le médicament B. Toutefois, sur le tableau de données individuelles, nous aurons l'information que les deux médicaments ont été prescrits en même temps et nous pourrons émettre une alerte de risque d'imputation à tort. Plus précisément, si le médicament B n'est pas responsable de l'effet, nous pouvons espérer que certains groupes d'individus n'auront pris que ce médicament et n'auront pas eu l'effet ; à l'opposé, certains individus pourront avoir pris uniquement le médicament A et avoir eu l'effet. Par conséquent, une analyse comparée des classes d'individus permettrait de voir plus facilement que le médicament B ne fait pas partie d'un signal.

Pour ce faire, nous avons simulé des données de pharmacovigilance selon le modèle de simulation numérique avec les paramètres suivants : $n_S = 300$, $q = 0.8$, $\varepsilon = 10^{-6}$, $(n, J, K) = (20\,000, 100, 400)$.

Nous avons ensuite supposé que 100 autres médicaments qui ne sont donc impliqués dans aucun signal, étaient prescrits avec les 100 premiers de la façon suivante. Nous avons associé les 50 premiers médicaments du premier groupe aux 50 premiers du second groupe. Puis les 25 autres médicaments du premier groupe sont associés ou coprescrits avec deux médicaments des 50 derniers du deuxième groupe. Pour les 25 derniers médicaments du premier groupe, il n'existe aucune coprescription avec les médicaments du second groupe. Une représentation schématique de ce protocole est présentée en figure 8.7. La probabilité de coprescription notée p_{co} a été fixée à 0.9, probabilité forte d'avoir des coprescription de médicaments. Cela signifie que lorsque 10 individus prennent le premier médicament, en moyenne, 9 individus prennent également le médicament coprescrit.

Ainsi, sur les 200 médicaments, seuls les 100 premiers peuvent être impliqués dans un signal, les 100 autres sont juste des médicaments coprescrits avec les médicaments responsables d'un effet indésirable. Au final, nous avons les paramètres suivants, $(n, J, K) = (20\,000, 200, 400)$ et nous

pouvons différencier deux groupes de médicaments : les médicaments pouvant être impliqués dans un signal et les médicaments coprescrits et non impliqués dans un signal.

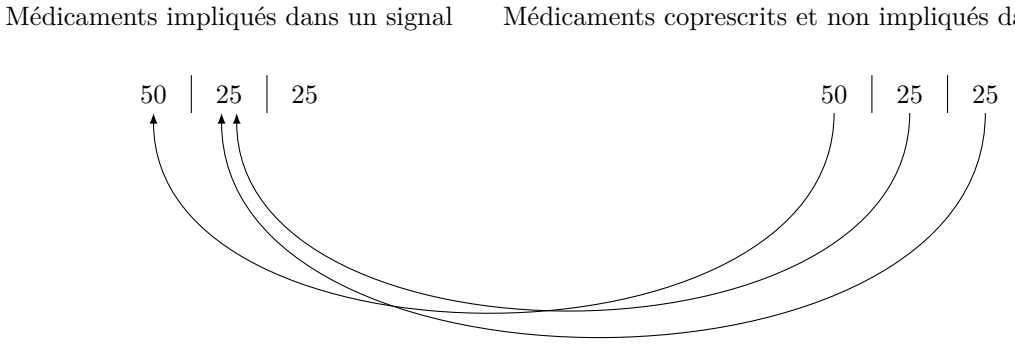


Figure 8.7 – Représentation schématique de l'effet de coprescription simulé.

Nous avons alors comparé chacune des méthodes présentées en section 8.1.2.a. en gardant l'ensemble des médicaments d'une part (expérience avec coprescription) et en retirant de l'étude les 100 médicaments de coprescription (expérience sans coprescription). Même si ces deux expériences ne se font pas sur le même nombre de médicaments (200 médicaments pour la première et 100 pour la deuxième), ce facteur n'est pas décisif dans l'interprétation des résultats car nous avons vu dans les exemples précédents que même si le nombre de médicaments est plus élevé et donc plus difficile à traiter, les méthodes gardent également de bonnes performances.

Les courbes ROC pour chacune des méthodes et pour ces deux expériences sont présentées en figures 8.8, 8.9 et 8.10. Nous remarquons que pour l'expérience avec coprescription, à taux de vrais positifs fixé, le nombre de faux positifs est plus élevé que celui de l'expérience sans coprescription, et ce pour toutes les méthodes. Ainsi, après vérification, les médicaments coprescrits induisent des faux positifs et ce très rapidement pour chacune des méthodes étudiées.

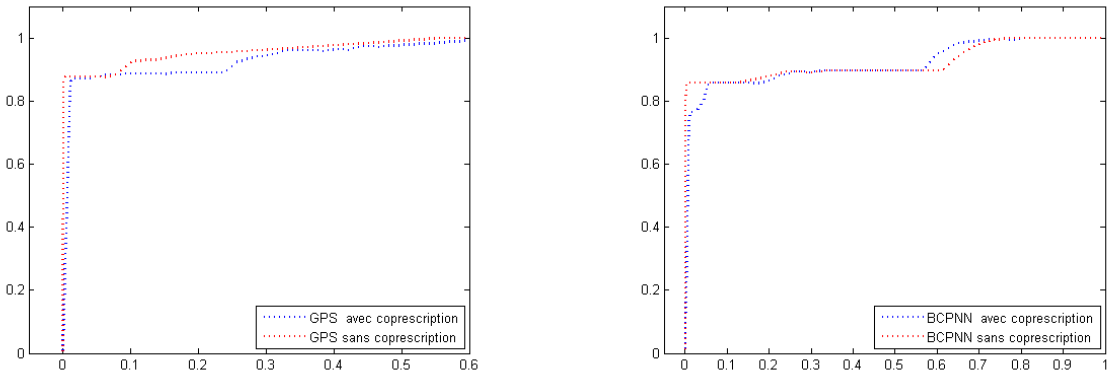


Figure 8.8 – Courbes ROC sans coprescription (en bleu) et avec coprescription (en rouge) de la méthode GPS (à gauche) et de la méthode BCPNN (à droite).

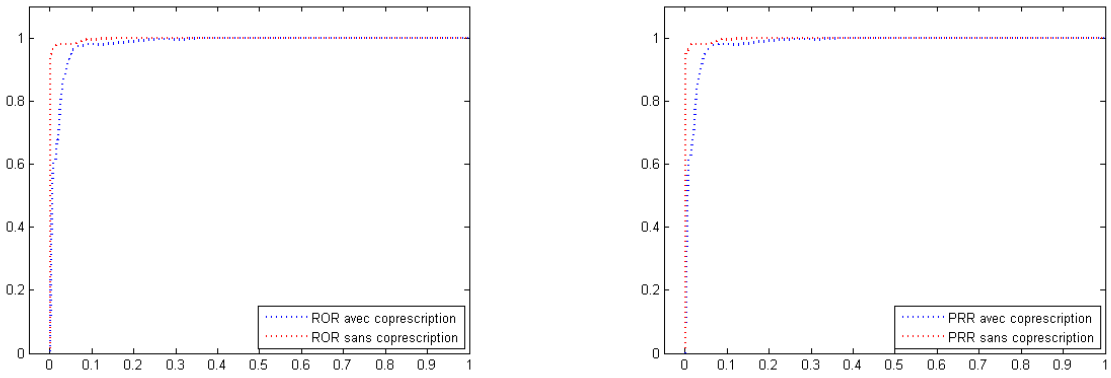


Figure 8.9 – Courbes ROC sans coprescription (en bleu) et avec coprescription (en rouge) de la méthode ROR (à gauche) et de la méthode PRR (à droite).

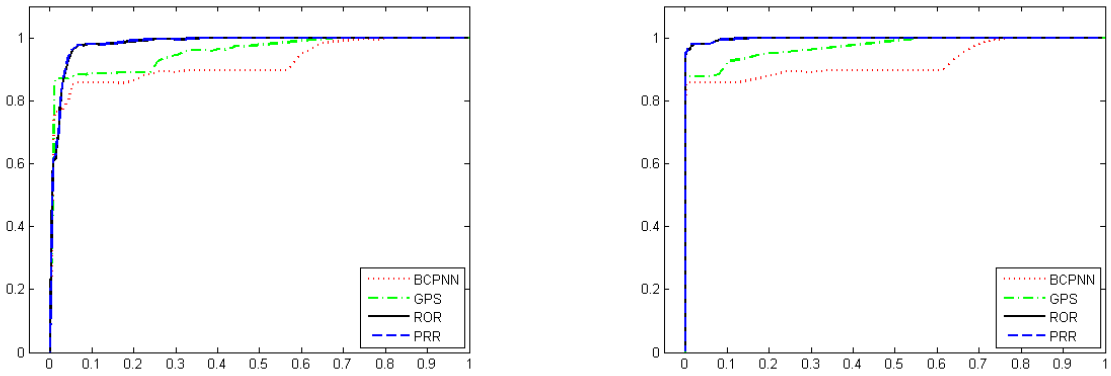


Figure 8.10 – Courbes ROC de toutes les méthodes sans coprescription (à droite) et avec coprescription (à gauche).

Dans cette section, nous avons ainsi montré que les méthodes de détection de signaux sur les données de contingence ne permettent pas de pallier le phénomène de coprescription qui peut alors entraîner une augmentation de faux positifs.

Intérêt de la méthode proposée face à la problématique de la coprescription

Nous étudions ici le comportement de la méthodologie proposée en section 8.1.2 face au phénomène de coprescription, en utilisant les mêmes données simulées du paragraphe précédent qui ont permis de tester les méthodes usuelles vis-à-vis de ce même phénomène. Rappelons que les paramètres utilisés sont les suivants : $n_S = 300$, $q = 0.8$, $\varepsilon = 10^{-6}$, $(n, J, K) = (20\,000, 200, 400)$, $p_{\text{co}} = 0.9$.

Nous avons alors effectué une classification croisée du tableau de contingence. Le critère ICL a sélectionné (29,22) classes en ligne et colonne. De même, nous notons que beaucoup de signaux ont été regroupés ensemble dans des blocs. De plus les blocs possédant les plus fortes intensités γ sont les blocs englobant les signaux générés (voir figure 8.11 et tableau 8.4).

Puis nous avons utilisé la procédure *Bi-KM1* étendue à partir des classes obtenues pour les lignes et colonnes du tableau de contingence, et d'une partition initiale de 2 classes en lignes obtenue grâce

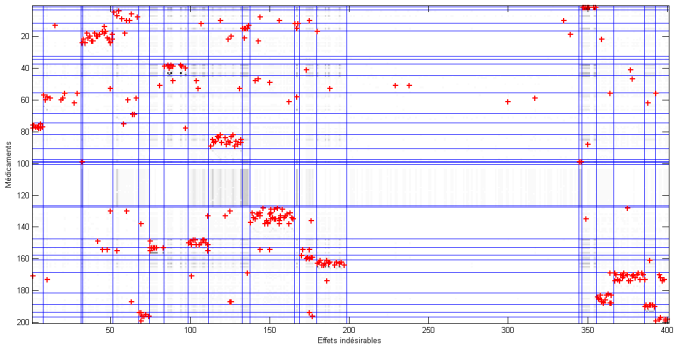


Figure 8.11 – Tableau de contingence réorganisé à l'aide des partitions estimées. Les croix rouges représentent les signaux.

γ	numéro classe en ligne	numéro classe en colonne
3,72e-05	29	22
3,53e-05	12	1
2,97e-05	7	6
2,11e-05	25	20
1,50e-05	5	4
1,37e-05	11	16
1,27e-05	6	14
1,19e-05	11	5
9,55e-06	26	19
9,14e-06	1	17

Table 8.4 – Les 10 blocs avec les intensités γ les plus fortes rangés par ordre décroissant.

à l'algorithme des *k-means*. Le critère ICL a sélectionné le triplet $(30, 33, 23)$.

D'une part, quand nous examinons les 50 premiers couples médicaments/coprescrits (voir figure 8.7), nous remarquons que les médicaments de 46 couples (soit 92%) sont dans des classes séparées, ce qui signifie que l'algorithme a effectué une distinction entre le médicament impliqué dans un signal et le médicament coprescrit non impliqué dans un signal pour chacun des 46 couples.

D'autre part, sur la figure 8.12, intéressons-nous aux classes d'individus 16 et 17 (des nombres verts ont été ajoutés pour faciliter la lecture). Nous pouvons constater que les blocs (16,16), (17,2) et (17,16) de la matrice x des individus/médicaments sont d'intensités très fortes. De même, les blocs (16,3) et (17,3) de la matrice y des individus/effets sont d'intensités très fortes. Si nous ne regardons que la classe 17 des individus, nous pourrions supposer que les effets de la classe 3 sont autant dûs aux médicaments de la classe 2 qu'au médicament de la classe 16. Or, d'après la matrice S, le seul médicament responsable de tous les effets indésirables de la classe 3, est celui de la classe 16. L'effet de coprescription peut être ainsi résolu grâce à la classe 16 des individus qui ont pris uniquement le médicament de la classe 16 et ont eu les effets considérés. Nous pouvons alors imputer les effets indésirables de la classe 3 seulement au médicament de la classe 16.

Par le même raisonnement et de façon plus flagrante, nous pouvons nous intéresser aux classes d'individus 1, 5, 14 et 30 (en rouge) ainsi que les classes de médicaments 15, 18, 19 et 30. Grâce à la matrice de signaux S, nous savons que les effets indésirables de la classe 11 sont tous induits par le médicament de la classe 30. Mais comme les médicaments des classes 15, 18 et 19 sont très

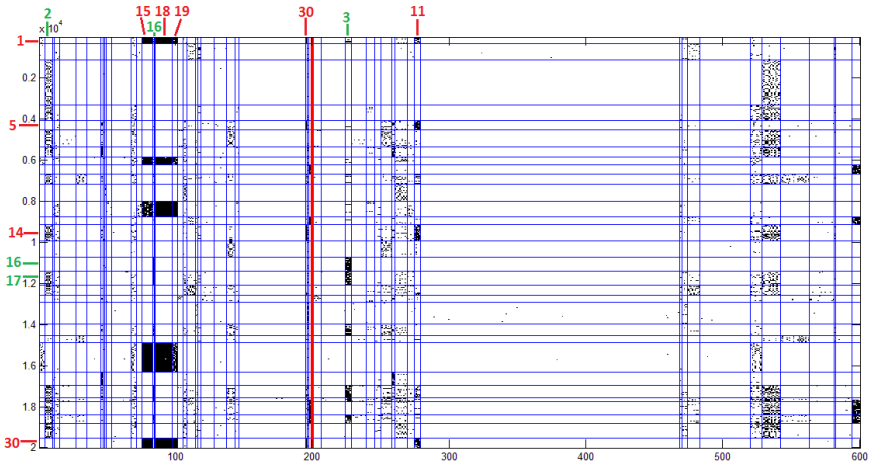


Figure 8.12 – Matrices x et y simulées et réorganisées à l'aide des partitions estimées.

souvent prescrits en même temps que le médicament de la classe 30 (au moins pour les classes d'individus 1 et 30), le tableau de contingence notifie fortement à tort ces couples. À l'opposé, la présence des classes d'individus 5 et 14 obtenues par notre méthodologie, permet de ne pas imputer aux médicaments des classes 15, 18 et 19, les effets indésirables de la classe 11.

Nous voyons ainsi qu'une analyse croisée des classes des individus apporte un éclairage nouveau à la détection de signaux et permet notamment de pallier le phénomène de coprescription.

8.3 Application aux données individuelles réelles de pharmacovigilance

8.3.1 Présentation

Les données utilisées pour cette application correspondent aux **219 340** notifications spontanées, collectées par le système de pharmacovigilance français entre 2000 et 2010 (Marbac et al. (2016)). Ces données initialement codées au niveau *PT* (*Preferred term*, niveau de précision moyenne) dans la classification ATC (voir section 1.1.2.c), impliquent **2142** médicaments et **4216** événements indésirables qui ont donc été notifiés au moins une fois. Elles ne sont pas librement accessibles mais ont été rendues disponibles par l'équipe B2PHI (UMR 1181, INSERM, Villejuif) que nous remercions chaleureusement.

Par ailleurs, il existe d'autres bases plus conséquentes que la base française : la base américaine coordonnée par la Food and Drug Administration (FDA) et la base de l'Organisation Mondiale de la Santé administrée par l'Uppsala Monitoring Center en Suède, qui, en décembre 2004, contenaient respectivement environ 2.6 et 3.7 millions de notifications (Almenoff et al. (2007)).

Le tableau suivant résume toutes les informations relatives aux matrices x représentant la prise de médicaments de ces 219 340 individus et y représentant la survenue d'effets indésirables pour ces mêmes individus.

	x	y
Taille	$219340 \times 2142 = 469\ 826\ 280$	$219340 \times 4216 = 924\ 737\ 440$
Nombre de cellules non nulles	616 552 (0.13%)	382 081 (0.041%)

8.3.2 Comparaison des méthodes usuelles appliquées sur tableau de contingence avec une méthode naïve sur les données individuelles

Avant d'appliquer la théorie développée dans cette thèse sur les données réelles, nous proposons une méthode naïve basée sur les données individuelles que nous comparons avec les méthodes usuelles sur le tableau de contingence pour montrer l'intérêt de travailler avec les données individuelles. Le principe de cette méthode consiste simplement à estimer de façon empirique la probabilité d'avoir un effet indésirable sachant que nous avons pris un médicament. Plus précisément, pour tout couple (j,k) de médicament j et d'effet k, nous recherchons la probabilité :

$$p(\text{avoir effet}_k|\text{avoir pris médicament}_j) = \frac{p(\text{avoir effet}_k \text{ et avoir pris médicament}_j)}{p(\text{avoir pris médicament}_j)}.$$

Pour cela et au vu du grand nombre d'individus que nous avons, nous faisons une estimation empirique à l'aide des tableaux x des médicaments et y des effets. Nous calculons alors pour tout couple (j,k) la probabilité suivante

$$PE(\text{effet}_k|\text{médicament}_j) = \frac{\sum_{i=1}^{n} x_{ij}y_{ik}}{\sum_{i=1}^{n} x_{ij}} = \frac{c_{jk}}{n_{\cdot j}}.$$

Pour décider si un effet est causé par un médicament ou non, il suffit de choisir un seuil à partir duquel déclencher une alerte.

Nous avons comparé cette procédure naïve aux méthodes *classiques* présentées précédemment. Pour cela, nous avons utilisé le jeu de référence OMOP proposé par Ryan et al. (2013) contenant des témoins positifs et des témoins négatifs pour quatre effets indésirables d'intérêt : lésions hépatiques aigües (ALI), lésions rénales aigües (AKI), infection aigüe du myocarde (AMI), et saignements gastro-intestinaux supérieurs (GIB). Dans cet ensemble de référence, 399 couples témoins sont disponibles et parmi eux, 165 sont des témoins positifs et 234 sont des témoins négatifs. Ryan et al. (2013) précisent que la plupart des témoins positifs pour les effets AKI et GIB sont issus de résultats d'essais cliniques randomisés alors que la majorité des témoins positifs pour AMI et ALI sont basés sur des études de cas publiés. Notons que, même si ce jeu de données fournit un très grand nombre de témoins à la fois positifs et négatifs, il n'est absolument pas exhaustif et nous ne jugeons de la qualité des méthodes que sur certains effets indésirables. Sur la base étudiée ici, seuls 198 témoins y sont présents.

Comme toutes les méthodes étudiées ici, nécessitent un seuil, nous avons opté pour des comparaisons à l'aide de courbes ROC. Nous avons représenté les courbes ROC sur la figure 8.13 pour lesquelles nous avons également calculé l'aire sous la courbe (AUC) ; plus cette valeur est proche de 1, meilleure est la méthode. Bien que naïve, la méthode PE proposée donne le meilleur AUC de toutes les méthodes étudiées et est donc comparable à ces dernières.

Ce petit exemple montre l'intérêt d'étudier les données individuelles plutôt que le tableau de contingence.

Remarque. Notons qu'un intérêt non négligeable de cette méthode est de pouvoir être utilisée sur un très grand jeu de données rapidement. En effet, le calcul informatique consiste simplement à faire la somme sur toutes les lignes pour chaque colonne (donc une complexité en $\mathcal{O}(nJ)$) pour le dénominateur et d'utiliser le tableau de contingence pour le numérateur. La complexité finale est donc $\mathcal{O}(\max(nJ, JK))$ si on suppose avoir déjà le tableau de contingence et de $\mathcal{O}(nJL)$ s'il faut recalculer ce dernier.
En particulier, les moyens informatiques mis à notre disposition ne nous permettaient pas d'ouvrir les deux matrices (respectivement de tailles 3.5 Gb et 7 Gb) en même temps mais nous avons pu quand même faire les calculs.

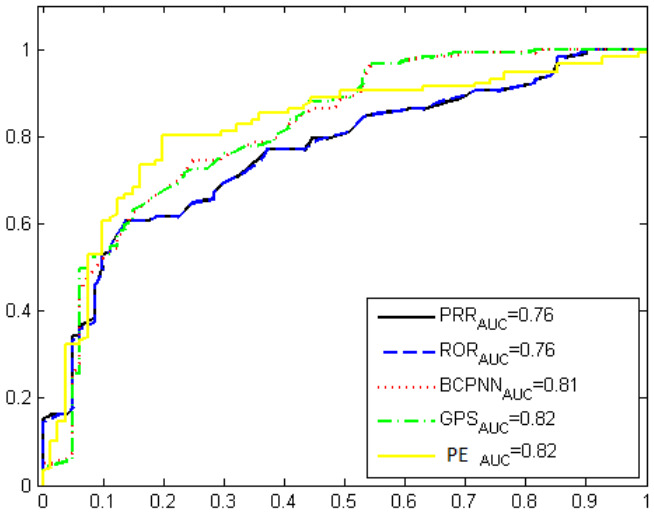


Figure 8.13 – Courbes ROC permettant la comparaison des méthodes existantes sur le tableau de contingence de pharmacovigilance.

8.3.3 Application au jeu de données réelles : base française de pharmacovigilance entre 2002 et 2010

Pour conclure cette thèse, nous avons souhaité appliquer la méthodologie proposée sur le jeu de données réelles :

- classification croisée du tableau de contingence via le modèle LBM Poisson normalisé,
- classification croisée des données individuelles via le modèle $MLBM$ en utilisant l'information a priori issue de la classification croisée du tableau de contingence.

Le principal problème que nous avons rencontré est d'ordre logistique : le matériel informatique mis à notre disposition ne nous permettait pas de lire dans un format usuel les deux matrices x et y en même temps.

Pour contourner le manque de mémoire vive, nous avons choisi de ne prendre de façon aléatoire qu'un individu sur 20, soit environ 10 000 individus.

Nous avons alors effectué une classification croisée du tableau de contingence construit en utilisant ces seuls individus, via le modèle LBM Poisson normalisé.

Vu que nous avons réduit le nombre d'individus, seuls 89 signaux de référence sur les 198 recensés de l'OMOP sont notifiés au moins une fois dans notre nouvelle base (voir tableau 8.5).

Effet indésirable	AMI	GIB	ALI	AKI
Témoin positif	4	9	44	12
Témoin négatif	3	10	3	4

Table 8.5 – Nombre de signaux de l'OMOP présents (avec au moins une notification) dans la base française de pharmacovigilance entre 2002 et 2010 restreinte aux 10 000 individus choisis aléatoirement.

Le critère ICL a sélectionné (10,13) classes en ligne et colonne, ce nombre de classes étant logiquement inférieur à celui obtenu sur le tableau de contingence complet du chapitre 5. Puis nous

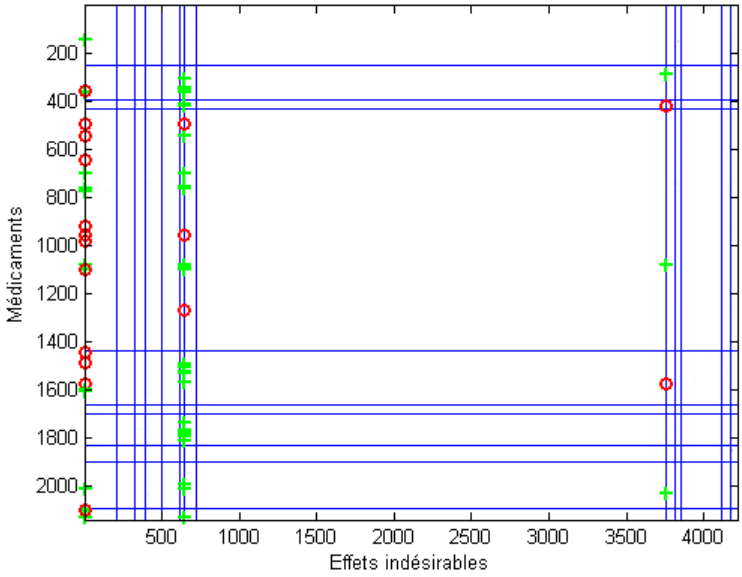


Figure 8.14 – Tableau de contingence construit sur les 10 000 individus choisis aléatoirement et réorganisé à l'aide des partitions estimées. Les croix représentent les témoins positifs de l'OMOP et les ronds les témoins négatifs.

avons utilisé la procédure *Bi-KM1* étendue à partir des classes obtenues pour les colonnes et d'une partition initiale de 2 classes en lignes obtenue grâce à l'algorithme des *k-means*. Le critère ICL a sélectionné le triplet $(3, 11, 13)$.

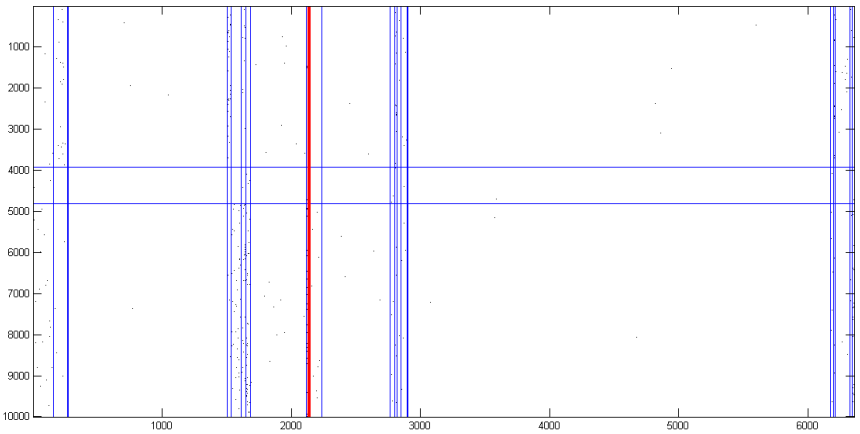


Figure 8.15 – Matrices x et y restreintes aux 10 000 individus choisis aléatoirement et réorganisées à l'aide des partitions estimées.

Nous pouvons remarquer sur la figure que l'effet indésirable associé au plus de vrais positifs (AKI) se retrouve dans une petite classe.

Pour les tableau des données individuelles, nous pouvons remarquer que peu de classes en lignes ont été formées. De plus, la plupart des classes des médicaments et des effets sont très petites.

Comme pour les données simulées, nous avons représenté le tableau 8.6 avec les intensités α et β les plus grandes pour chaque classe d'individus. Nous pouvons remarquer que la classe $g = 3$ possède 892 individus et que les classes de médicaments et d'effets associés contiennent 5 témoins positifs de l'OMOP. De même, la classe $g = 2$ possède 2 signaux positifs pour les classes de médicaments et d'effets associés.

De plus, les médicaments impliqués dans un signal et présents dans la classe 3 sont des inhibiteurs de la transcriptase inverse nucléosides (code ATC J05AF). Notre procédure a donc rassemblé des individus prenant la même classe thérapeutique de médicaments entraînant l'effet indésirable ALI.

n^o classe d'individus	plus grande valeur de α	plus grande valeur de β	classe de médicaments correspondante	classe d'effets correspondante	effectif des individus	nombre de médicaments	nombre d'effets	nombre de signaux
2	0,23	0,03	3	10	892	8	9	5
3	0,08	0,02	10	13	5193	8	13	0
1	0,02	0,05	5	10	3915	28	9	2

Table 8.6 – Groupes d'individus rangés par ordre décroissant des valeurs de α les plus fortes pour chaque groupe d'individus.

En augmentant la capacité de mémoire vive nécessaire à l'analyse des tableaux x et y présentant les 219 340 individus, nous aurions pu traiter le jeu de données réelles en entier, à partir de la partition du tableau de contingence obtenue dans le chapitre 5. Notons cependant que l'analyse sur les 10 000 individus a permis de tester la méthode et de valider la pertinence de la méthodologie.

9

Conclusion et perspectives

9.1 Conclusion

Durant cette thèse, nous avons proposé une méthodologie pour traiter à la fois le tableau de contingence et les données individuelles de grandes dimensions en pharmacovigilance. En effet, une application de la classification croisée à ces données, approche novatrice en pharmacovigilance, nous a conduit à effectuer un certain nombre d'avancées dans l'étude du modèle des blocs latents d'un point de vue théorique et algorithmique.

Dans les avancées de cette thèse, nous pouvons notamment retenir :

Indice de classification croisée. Nous avons proposé un indice de classification croisée nommé CARI, inspiré du consensuel indice *Adjusted Rand Index*. Contrairement à des extensions d'indice de classification simple qui ne prennent pas pleinement en compte la nature spécifique des données et des modèles en classification croisée (Wyse et al. (2016)), l'indice que nous avons développé a été réfléchi d'un point de vue de coclustering. De plus, contrairement à l'indice développé par Lomet (2012), l'indice CARI est calculable en pratique pour des partitions présentant un grand nombre de classes, ce qui est parfaitement adéquat dans le cadre de données massives.

Sélection de modèles : aspects théoriques et algorithmiques. Pour le *LBM* Poisson normalisé, nous avons développé deux critères de sélection de modèles et étudié leur comportement ainsi que leur sensibilité vis-à-vis des hyperparamètres du modèle.

D'un point de vue algorithmique, nous avons proposé une nouvelle procédure *Bi-KM1* qui permet de parcourir de manière non exhaustive le nombre de classes en ligne et en colonne lors de l'étape de la sélection de modèle. Celle-ci s'avère particulièrement utile lorsque nous devons explorer des

bases de données de grandes dimensions.

Extension proposée : Modèle des blocs latents multiple. Nous avons développé un nouveau modèle pour traiter les données individuelles de pharmacovigilance, inspiré du modèle des blocs latents. Celui-ci permet d'obtenir une classification croisée de deux tableaux binaires, en leur imposant le même classement en ligne. Pour ce modèle, nous avons alors étudié une procédure d'estimation de ses paramètres, énoncé des conditions suffisantes de son identifiabilité et établi des critères de sélection de modèles.

Un modèle de simulation numérique en pharmacovigilance. Nous avons proposé un modèle de simulation numérique des données individuelles de pharmacovigilance qui nous permet de confronter les méthodes existantes entre elles. De plus, la sensibilité des méthodes face au phénomène de coprescription a pu être étudiée et l'intérêt de la méthodologie proposée vis-à-vis de ce même phénomène a été également mis en exergue.

Méthodologie en pharmacovigilance. Nous avons proposé une méthodologie en pharmacovigilance qui consiste dans un premier temps à tirer parti de l'information présente dans le tableau de contingence de pharmacovigilance en y effectuant une classification croisée. Cette dernière fournit déjà des zones intéressantes plus réduites à explorer par des experts en pharmacovigilance. Ensuite, cette information est utilisée a priori pour traiter les données individuelles de manière à effectuer une classification croisée simultanée des matrices x et y représentant ces données. Cette méthode prend en compte pour la première fois en pharmacovigilance, l'interaction entre plusieurs effets indésirables. Elle peut permettre à terme de fournir des profils de personnes à risque pour les effets secondaires (enfants, personnes âgées, femmes enceintes...). De plus, les effets secondaires causés par la prise de deux médicaments joints pourront être également détectés.

9.2 Perspectives

Ces constatations nous amènent à proposer de nombreuses nouvelles pistes à développer :

Indices de classification croisée. L'indice que nous avons proposé durant cette thèse semble avoir un comportement similaire à l'indice de classification croisée développé dans Lomet (2012) mais ce dernier n'est pas calculable en pratique lorsque le nombre de classes dépasse 8. Par ailleurs, il serait intéressant d'étendre cet indice dans le cadre du $MLBM$. Ceci reviendrait essentiellement à inclure la même partition des lignes pour les deux matrices.

Étude de l'échantillonneur de Gibbs dans le cas de matrices creuses binaires. Durant nos travaux, nous nous sommes aperçus des limites des choix des hyperparamètres de l'échantillonneur de Gibbs proposés par Keribin et al. (2015) dans le cadre de matrices creuses. En effet, il semblerait qu'une loi uniforme ait un effet négatif sur l'affectation a posteriori des classes. Il serait intéressant de proposer une loi plus adaptée dans ce cadre, par exemple une loi Beta(a, b) avec $a \gg b$ et un choix approprié des hyperparamètres. Une autre piste potentielle serait de mettre également des lois a priori sur les hyperparamètres.

Extension du modèle des blocs latents multiple. Le modèle des blocs latents multiple développé dans le chapitre 6 peut être appliqué à d'autres contextes que la pharmacovigilance dès que l'on souhaite effectuer une classification d'individus via des variables qui peuvent se diviser en deux catégories : variables réponses et variables explicatives. De plus, nous pouvons l'étendre aux cas où les données à traiter se représenteraient sous la forme de plus de deux tableaux. À ce sujet des développements récents autour ce modèle ont permis d'étudier des données écologiques (**?**). Par ailleurs, ce modèle pourrait permettre l'étude de données mixtes et nous pourrions imaginer des variables de différents types (catégorielles, ordinales, etc...).

Adaptation de la procédure d'estimation à MapReduce. Comme nous travaillons avec des bases de données de grandes dimensions, il serait profitable d'étudier si les algorithmes d'estimation peuvent être adaptés à ce patron d'architecture informatique, ce qui permettrait de traiter plus efficacement et rapidement les données réelles. De plus, une piste intéressante pour améliorer les méthodes existantes, telles que la régression logistique, serait d'utiliser les développements récents autour du "sketching" proposés notamment par Richardson *et al.*.

Amélioration du modèle de simulation numérique. Le modèle de simulation numérique proposé peut être enrichi en introduisant des interactions entre effets indésirables, des effets de masquage et coprescription plus complexes, ou l'introduction de catégories de population particulières par exemple.

Analyse des résultats obtenus en collaboration avec un pharmacologue. Pour les données réelles, la prochaine piste à étudier est l'utilisation de nos méthodes sur les tableaux complets des données individuelles. Une fois les classes obtenues, il serait important d'une part d'avoir les profils des individus afin de mieux comprendre les classes formées et d'autre part d'avoir le retour des pharmacologues sur les classes (et donc potentiels signaux) formées.

Bibliographie

I. Ahmed. *Détection automatique de signaux en pharmacovigilance: Approche statistique fondée sur les comparaisons multiples*. Thèse, Université Paris Sud, 2009.

I. Ahmed, F. Thiessard, G. Miremont-Salame, B. Begaud, and P. Tubert-Bitter. Pharmacovigilance data mining with methods based on false discovery rates: a comparative simulation study. *Clinical Pharmacology & Therapeutics*, 88(4), 2010.

I. Ahmed, A. Pariente, and P. Tubert-Bitter. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*, page 0962280216643116, 2016.

J. Almenoff, E. Pattishall, T. Gibbs, W. DuMouchel, S. Evans, and N. Yuen. Novel statistical tools for monitoring the safety of marketed drugs. *Clinical Pharmacology & Therapeutics*, 82(2): 157–166, 2007.

J. Aubert, T. Ha, and T. MaryHuard. Modele à blocs latents pour l'analyse de données métagénomiques. In $46^{ème}$ *journées de Statistiques de la SFdS*, 2014.

A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998.

J.-P. Baudry and G. Celeux. Em for mixtures. *Statistics and Computing*, 25(4):713–726, 2015.

Y. Ben Slimen, S. Allio, and J. Jacques. Model-based co-clustering for functional data. In *48e Journées de Statistique, SFdS*, Montpellier, France, June 2016. URL `http://papersjds16.sfds.asso.fr/submission_111.pdf`.

C. Biernacki and J. Jacques. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, pages 1–15, 2015.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

V. Brault. *Éstimation et sélection de modèle pour le modèle des blocs latents*. PhD thesis, Université Paris Sud 11, 2014.

V. Brault, C. Keribin, and M. Mariadassou. Consistency and asymptotic normality for the maximum likelihood estimator in the latent block model. In *48e Journées de Statistique, SFdS*, Montpellier, France, June 2016. URL `http://papersjds16.sfds.asso.fr/submission_140.pdf`.

E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, 1999.

O. Caster, G. N. Norén, D. Madigan, and A. Bate. Large-scale regression-based pattern discovery: The example of screening the who global drug safety database. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(4):197–208, 2010.

A. Channarond, J.-J. Daudin, S. Robin, et al. Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601, 2012.

A. P. Dempster, N. M. Laird, D. B. Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

W. DuMouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999.

W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item associations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 67–76. ACM, 2001.

S. Evans, P. C. Waller, and S. Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, 10(6): 483–486, 2001.

S. Frühwirth-Schnatter. *Mixtures : estimation and applications*. Wiley, 2011. ISBN 9781119993896 111999389.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

G. Govaert and M. Nadif. Clustering of contingency table and mixture model. *European Journal of Operational Research*, 183:1055–1066, 2007.

G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52:3233–3245, 2008.

G. Govaert and M. Nadif. *Co-Clustering*. ISTE Ltd and John Wiley & Sons, Inc, 2013.

R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, and N. H. Shah. Performance of pharmacovigilance signal-detection algorithms for the fda adverse event reporting system. *Clinical Pharmacology & Therapeutics*, 93(6):539–546, 2013.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Series A*, 62:49–66, 2000.

C. Keribin, G. Govaert, and G. Celeux. Estimation d'un modèle à blocs latent par l'algorithme SEM. In *42e Journées de Statistique, SFdS*, Marseille, France, May 2010. URL http://hal.archives-ouvertes.fr/hal-00554409/en/.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

E. Lebarbier and T. Mary-Huard. Le critère BIC : fondements théoriques et interprétation. Rapport de recherche RR-5315, INRIA, 2004. URL http://hal.inria.fr/inria-00070685/en/.

A. Lomet. *Sélection de modèle pour la classification croisée de données continues*. Thèse, Université de Technologie de Compiègne, Décembre 2012.

A. Lomet, G. Govaert, and Y. Grandvalet. Un protocole de simulation de données pour la classification croisée. In *44eJournées de Statistique de la SFdS*, 2012.

G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1-2):303–324, 2016.

M. Marbac, P. Tubert-Bitter, and M. Sedki. Bayesian model selection in logistic regression for the detection of adverse drug reactions. *Biometrical Journal*, 58(6):1376–1389, 2016.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2008.

G. Miller and H. Britt. A new drug classification for computer systems: the atc extension code. *International Journal of Bio-Medical Computing*, 40(2):121–124, 1995.

G. N. Norén, A. Bate, R. Orre, and I. R. Edwards. Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine*, 25(21):3740, 2006.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

V. Robert, G. Celeux, and C. Keribin. Un modèle statistique pour la pharmacovigilance. In *47èmes Journées de Statistique de la SFdS*, 2015.

V. Robert, G. Celeux, C. Keribin, and P. Tubert-Bitter. Modele des blocs latents et sélection de modeles en pharmacovigilance. In *48èmes Journées de Statistique de la SFdS*, 2016.

E. Roux, F. Thiessard, A. Fourrier, B. Begaud, and P. Tubert-Bitter. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Transactions on Information Technology in Biomedicine*, 9(4):518–527, 2005.

P. B. Ryan, M. J. Schuemie, E. Welebob, J. Duke, S. Valentine, and A. G. Hartzema. Defining a reference set to support methodological research in drug safety. *Drug safety*, 36(1):33–47, 2013.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. doi: 10.2307/2958889. URL http://dx.doi.org/10.2307/2958889.

G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

A. Szarfman, S. G. Machado, and R. T. O'neill. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fda's spontaneous reports database. *Drug Safety*, 25(6):381–392, 2002.

P. G. van der Heijden, E. P. van Puijenbroek, S. van Buuren, and J. W. van der Hofstede. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. *Statistics in Medicine*, 21(14):2027–2044, 2002.

E. P. van Puijenbroek, A. Bate, H. G. Leufkens, M. Lindquist, R. Orre, and A. C. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, 11(1):3–10, 2002.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

J. Wyse, P. Latouche, and N. Friel. Inferring structure in bipartite networks using the latent block model and exact ICL. *Network Science*, 2016.