De la classification à la classification croisée : une approche basée sur la modélisation

Christine Keribin

Université Paris-Sud

École doctorale de mathématiques Hadamard (ED 574)
Laboratoire de mathématique d'Orsay (UMR 8628 CNRS)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Christine KERIBIN

De la classification à la classification croisée : une approche basée sur la modélisation

Rapporteurs :	Gérard BIAU Florence FORBES Pascal MASSART

Date de soutenance : 26 novembre 2019

Composition du jury :	Gérard BIAU	(Rapporteur)
	Gilles BLANCHARD	(Président)
	Stéphane CANU	(Examinateur)
	Florence FORBES	(Rapportrice)
	Estelle KUHN	(Examinatrice)
	Béatrice LAURENT-BONNEAU	(Examinatrice)
	Pascal MASSART	(Rapporteur)
	Gilles CELEUX	(Invité)

Résumé Ce mémoire d'habilitation retrace des travaux portant principalement sur la classification non supervisée par modélisation probabiliste et sur la question connexe du choix de modèle. Après avoir rappelé l'apport des modèles de mélange à la classification non supervisée (*clustering*), le modèle des blocs latents (LBM), un modèle de mélange étendu à la classification simultanée (*co-clustering*) des lignes et des colonnes d'un tableau de données, est introduit. Des contributions théoriques (identifiabilité, consistance et normalité asymptotique des estimateurs) et méthodologiques (estimation par EM variationnel, EM stochastique, EM variationnel bayésien, échantillonneur de Gibbs, choix de modèle via le critère ICL) sont présentés. Le LBM est étendu au modèle de blocs latents multiples (MLBM) pour traiter des données individuelles en pharmacovigilance et un algorithme glouton de parcours des modèles est proposé. L'étude de données d'IRM fonctionnelle, pour lesquelles le nombre d'individus est très inférieur au nombre de variables, a permis d'explorer le problème de la grande dimension suivant deux directions : utilisation de l'inférence bayésienne à des fins de régularisation (modèle MSBR –Multi Sparse Bayesian Regression) ; réduction drastique de la dimension tout en gardant des résultats interprétables (*clustering* de variables contraintes spatialement *supervisé* par la prédiction de la cible). Enfin, quelques contributions dans des domaines plus éloignés de modélisation de données applicatives (génomiques, météorologiques, phylogénétiques ou financières) illustrent comment des besoins applicatifs font surgir des questions théoriques ou méthodologiques intéressantes.

Abstract This habilitation thesis retraces works focusing mainly on model based clustering and the related issue of model choice. After recalling the contribution of mixture models to the unsupervised framework of *clustering*, the latent block model (LBM), a mixture model extended to the simultaneous clustering (*co-clustering*) of rows and columns of a data table, is introduced. Theoretical (identifiability, consistency and asymptotic normality of the estimators) and methodological contributions (estimation by variational EM, stochastic EM, Bayesian variational EM, Gibbs sampler, model choice with the ICL criterion) are presented. The LBM is extended to the Multiple Latent Block Model (MLBM) to process individual data in pharmacovigilance and a greedy algorithm to scan the model set is proposed. The study of functional MRI data, for which the number of individuals is much smaller than the number of variables, made it possible to explore the large dimension paradigm in two directions : use of Bayesian inference as a regularization tool (MSBR model - Multi Sparse Bayesian Regression) ; drastic dimension reduction while keeping interpretable results (*clustering* of spatially constrained variables *supervised* by the prediction of the target). Finally, some contributions in less related domains (data modeling in genomics, meteorology, phylogenetics or finance) illustrate how applications bring up interesting theoretical or methodological issues.

Remerciements

La pensée des remerciements s'est imposée avant même d'avoir écrit une ligne du manuscrit, tant je me suis sentie portée et encouragée par l'ensemble de la communauté. Au moment de faire part de ma gratitude, je me rends compte de l'entreprise délicate de sa rédaction, et je prie par avance les personnes que j'aurais pu oublier de bien vouloir m'en excuser.

Je suis très reconnaissante à Pascal Massart, Gérard Biau et Florence Forbes d'avoir accepté de rapporter ce mémoire, pour l'intérêt qu'ils ont porté à mon travail et pour leur présence à mon jury. Leur avis m'est important.

Pascal, tu m'as ouvert la possibilité d'une voie universitaire en appuyant ma candidature au DEA en parallèle de mon travail d'ingénieur. Tu m'as renouvelé ta confiance en me proposant la responsabilité pour Paris-Sud du M1 Mathématiques appliquées, puis en m'encourageant dans la voie de l'habilitation. Merci d'avoir veillé sur ma carrière, merci pour tes conseils avisés.

Gérard, merci d'avoir pris de ton temps parmi toutes tes sollicitations. Merci pour tes conseils et ton soutien constants au long des années ainsi que pour nos nombreuses discussions au sujet de la SFdS.

Florence, j'ai eu plaisir à nos échanges enrichissants sur les algorithmes et la vision bayésienne.

Je remercie également chaleureusement Gilles Blanchard, Stéphane Canu, Estelle Kuhn, Béatrice Laurent-Bonneau et Gilles Celeux d'avoir accepté de participer à ma soutenance et d'y apporter leur expertise.

Parmi les membres du jury, Gilles Celeux tient une place particulière. Gilles, tu m'as remise sur les rails de la recherche à un moment où je m'en étais éloignée, tu m'as initiée à la direction de recherche, et tu n'as cessé de me pousser hors de mes retranchements lors de nos multiples collaborations. Je t'en suis particulièrement reconnaissante. Tu sais mettre le doigt sur les problèmes, et toujours avec bienveillance. C'est un réel plaisir de travailler avec toi, et je suis ravie de savoir que nous avons encore du pain sur la planche.

Je remercie Frédéric Paulin d'avoir coordonné ma candidature à l'habilitation avec attention et précision. Merci aux membres du comité interne dont Stéphane Robin que je n'ai pas encore cité. Merci à Jean-François Le Gall, David Harari, la CCSU de mathématiques et tou.te.s ceux.elles qui m'ont soutenue dans l'obtention d'une délégation CNRS, catalyseur puissant à la finalisation de mon habilitation.

Je remercie Élisabeth Gassiat d'avoir fondé le socle de ma recherche scientifique pendant ma thèse. Dans un autre registre, merci Elisabeth d'avoir piloté le projet de notre bel institut de mathématique.

Yves Rozenholc est le premier à m'avoir instillé l'idée que je puisse écrire et soutenir une HDR. Mille mercis Yves, ce travail n'aurait pas vu le jour sans ton insistance. Tu m'as redonné confiance dans les moments de doute. J'ai également apprécié la richesse et l'abondance de tes idées lors de notre collaboration. J'espère que nous aurons le plaisir de retravailler ensemble.

Cette habilitation doit beaucoup aux personnes avec lesquelles j'ai eu la chance de travailler. Parmi celles que je n'ai pas encore citées, je remercie sincèrement Gérard Govaert de m'avoir initiée à la classification croisée, Marie-Laure Martin-Magniette de m'avoir fait connaître les réseaux de gènes, Tatiana Popova de m'avoir montré une voie de définition des cartes d'identité tumorale, Bertrand Thirion de m'avoir fait découvrir l'imagerie par résonance magnétique fonctionnelle, Charles Hernandez et Philippe Drobinski pour leur partage des problématiques de feux de forêts, Jean-Marc Guillard pour m'avoir fait revivre le temps d'un projet notre ancienne collaboration à Dassault-Systèmes, et Mahendra Mariadassou toujours disponible pour discuter sur des finesses de preuves asymptotiques. Merci pour nos échanges passionnants, vous m'avez apporté une grande richesse.

Je remercie également chaleureusement Vincent Michel, Vincent Brault et Valérie Robert pour l'accueil qu'ils m'ont réservé à la co-direction de leur thèse. Vous avez accepté mon apprentissage de la direction avec une grande bienveillance. Merci en particulier à Vincent Brault, toujours plein d'idées, à l'écoute et prêt à rendre service, aussi bien pendant la thèse que lors de notre collaboration ultérieure. Merci à Valérie qui consacre maintenant une part de son temps libre à améliorer nos résultats.

Je suis reconnaissante de la confiance témoignée par mes nouveaux collaborateurs. Merci à Christophe Biernacki pour la richesse de nos discussions sur le (co)clustering et la direction commune de la thèse de Filippo Antonazzo. Merci à Gilles Stoltz de m'avoir proposé de codiriger la thèse de Rémi Coulaud et à Patrick Pamphile de m'accueillir dans celle d'Olivier Coudray. Je n'en remercie pas moins chaleureusement Filippo, Rémi et Olivier. Je suis consciente de ma responsabilité envers vous.

J'ai trouvé au laboratoire de mathématiques d'Orsay, et dans l'équipe de probabilité et statistique en particulier, un environnement très agréable et stimulant. J'ai toujours beaucoup de plaisir à y travailler. Merci à tous les collègues que j'ai pu y rencontrer aussi bien en recherche qu'en enseignement et à ceux.elles qui n'oublient jamais de venir me chercher pour aller déjeuner. Je remercie en particulier Marie-Anne Poursat de m'avoir montré les clés de l'enseignement en master. Merci pour nos discussions pédagogiques, pour ton soutien dans les bons et les mauvais jours. Merci à Liliane Bel et Yves Auffray entre autres d'avoir régulièrement pris des nouvelles de l'avancement de mon travail. Un grand merci aux gestionnaires et secrétaires qui nous rendent la vie administrative et le suivi pédagogique plus légers et à tous les membres du service informatique toujours prêt.e.s à nous aider.

D'autres portes m'ont également été ouvertes. Je remercie sincèrement Frédéric Jean et Eric Lunéville pour leur accueil à l'ENSTA permettant un M1 coopéré entre l'université et leur grand école, et Erwan Le Pennec de m'avoir fait participer à l'aventure de la formation continue en Data Science. Je remercie Jean-Michel Poggi de m'avoir amenée à la SFdS, me donnant l'occasion de faire, avec les Rendez-vous Méthodes et Logiciels, mes premières armes d'organisatrice de manifestations scientifiques. Merci à Valérie Girardin pour ses encouragements enthousiastes lors de nos déjeuners à refaire le monde des mathématiques.

Si j'ai été portée par le courant de la communauté universitaire, je n'aurais pas pu atteindre mon but sans le souffle de mon environnement familial. Merci à Marie-France, Patrick, Marie-Lou et Vincent de veiller sur moi. Merci au groupe lillois des marcheurs de Compostelle qui m'a laissé sortir mon ordinateur portable dans des endroits improbables. Une partie de ce manuscrit a été rédigée lors de l'ultime tronçon. Et merci infiniment à Philippe pour son soutien inconditionnel et son accompagnement sans faille tout au long des années, ainsi qu'à Loïc, Muriel et Yann qui illuminent notre vie.

Liste des travaux publiés

Articles soumis pour publication

1. Brault V., **Keribin C.**, Mariadassou M.: *Consistency and Asymptotic Normality of Latent Block Model Estimators*, https://arxiv.org/abs/1704.06629v2

Revues internationales à comité de relecture

2. Hernandez C., **Keribin C.**, Drobinski P., Turquety S.: *Statistical modelling of wildfire size and intensity: a step toward meteorological forecasting of summer extreme fire risk*, Annales Geophysicae, 33 (1495–1506), DOI=10.5194/angeo-33-1495-2015, **2015**

3. **Keribin C**., Brault V., Celeux G., Govaert G.: *Estimation and Selection for the Latent Block Model on Categorical Data.* Statistics and Computing 25(6), pp 1201–1216, **2015**

4. Michel V., Gramfort A., Varoqueux G., Eger E, **Keribin C**. and Thirion B.: *A supervised clustering approach for fmri-based inference of brain states*. Pattern Recognition - Special Issue on Brain Decoding, 45(6): 2041-2042, **2012**

5. Michel V., Eger E., **Keribin C**. and Thirion B.: *Multi-Class Sparse Bayesian Regression for fMRI-based prediction*, International Journal of Biomedical Imaging, IJBI/350838, **2011** http://www.hindawi.com/journals/ijbi/2011/350838/

Revues nationales à comité de relecture

6. **Keribin C**.: *A note on BIC and the slope heuristic*, Journal de la SFdS, Vol. 160, N°3, pp 133-139, **2019**

7. **Keribin C**., Liu Y., Popova T., Rozenholc Y.: *A mixture model to characterize genomic alterations of tumors*. Journal de la SFdS, Vol. 160, N°1, pp 130-148, **2019**

8. **Keribin C**. : *Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie*. Journal de la SFdS, vol 151, N°2, **2010**

Conférences internationales avec actes publiés

9. **Keribin C.**: *Some asymptotic properties of model selection criteria in the latent block model*, in 12th Scientific meeting CLADAG 2019, Cassino (Italie), **2019**

10. **Keribin C.**, Celeux G., Robert V. : *The Latent Block Model : a useful model for high dimensional data*, ISI 2017, 61st World Statistics Congress, Marrakech, Maroc, **2017**

11. **Keribin C.**, Brault V., Celeux G., Govaert G.: *Model selection for the binary latent block model*. in Proceedings of COMPSTAT 2012, Limassol, Cyprus, **2012**

12. Michel V., Eger E., **Keribin C**., Thirion B.: *Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis*. in International Workshop on Machine Learning in Medical Imaging (MLMI) In conjunction with MICCAI 2010, **2010**

13. Michel V., Eger E., **Keribin C**., Poline J.-B., Thirion B.: *A supervised clustering approach for extracting predictive information from brain activation images*. In IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA10) - IEEE Conference on Computer Vision and Pattern Recognition, **2010**

14. Michel V., Eger E., **Keribin C**. and Thirion B.: *Adaptive multi-class Bayesian sparse regression - an application to brain activity classification*. In MICCAI'09 Workshop on Analysis of Functional Medical Images, **2009**.

Conférences nationales avec actes publiés

15. **Keribin C.**, Biernacki C. : *Le modèle des blocs latents, une méthode régularisée pour la classification en grande dimension*, in 51èmes Journées de Statistique, Nancy, **2019**

16. Brault V., **Keribin C.**, Mariadassou, M. : *Équivalence asymptotique des vraisemblances observée et complète dans le modèle de blocs latents*, XXIV èmes Rencontres de la Société Francophone de Classification, Lyon, **2017**

17. Brault V., **Keribin C.**, Mariadassou, M. : *Normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de blocs latents*, in 48èmes Journées de Statistique, Montpellier, **2016**

18. Robert V., Celeux G., **Keribin C.** : *Modèle des blocs latents et sélection de modèles en pharmacovigilance*, in 48èmes Journées de Statistique, Montpellier, **2016**

19. Robert V., Celeux G., **Keribin C.** : *Un modèle statistique pour la pharmacovigilance*, in 47èmes Journées de Statistique, Lille, **2015**

20. Liu Y., **Keribin C.**, Popova T., Rozenholc Y. : *Statistical Estimation of Genomic Tumoral Alterations*, in 47èmes Journées de Statistique, Lille, **2015**

21. Brault V., Celeux G., **Keribin C**. : *Mise en oeuvre de l'échantillonneur de Gibbs pour le modèle des blocs latents*. in 46èmes Journées de Statistique, Rennes, **2014**

22. Brault V., Celeux G., **Keribin C**. : *Régularisation bayésienne du modèle des blocs latents*. in 44èmes Journées de Statistique, Bruxelles, Belgique, **2012**

23. **Keribin C**., Celeux G., Govaert G. : *Estimation d'un modèle à blocs latents par l'algorithme SEM.* in 42èmes Journées de Statistique, Marseille, France, **2010**

Rapports de recherche

24. **Keribin C**., Brault V., Celeux G., Govaert G.: *Estimation and Selection for the Latent Block Model on Categorical Data*. Rapport de recherché Inria RR-8264, **2013**

25. **Keribin C**. : *Les méthodes bayésiennes variationnelles et leur application en neuroimagerie : une étude de l'existant*. Rapport de recherche Inria n°7091, **2009**

Publications issues de la thèse

26. **Keribin C**., Haughton D.: *Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families*. Communications in Statistics, Theory and Methods, Vol. 32, N°7, pp 1373-1404, **2003**

27. Gassiat E., **Keribin C**.: *The Likelihood Ratio Test for the number of components of a mixture with Markov regime*. ESAIM: PS, Vol. 4, p. 25-52, **2000**

28. **Keribin C**.: *Consistent Estimation of the Order of Mixture Models*. Sankhya Series A, volume 62, Part. 1, pp 49-66, **2000**

29. **Keribin C.** : Estimation consistante de l'ordre de modèles de mélange. CRAS Série I, 326(2), **1998**

Table des matières

Introduction

Ce document restitue la synthèse de mes travaux de recherche. Mon parcours présente l'originalité d'une double expérience en entreprise d'une part, puis dans l'enseignement supérieur et la recherche d'autre part.

La première partie de ma carrière s'est déroulée à Dassault-Systèmes [1], actuel leader des éditeurs de conception et d'ingénierie digitales où j'ai débuté dans un poste d'ingénieur de développement en Conception Assistée par Ordinateur (CAO) en 3D. L'ajout de fonctionnalités dans le logiciel de maillage pour les éléments finis, puis la définition et la réalisation du logiciel de visualisation des résultats de calcul par éléments finis m'ont permis de faire mes premières armes industrielles, à l'interface des mathématiques et de l'informatique. J'ai ensuite géré des projets et dirigé des équipes comptant jusqu'à une dizaine d'ingénieurs. Au sein du département *Conception de Pièces Mécaniques*, le service *Modélisation Paramétrique et Feature* dont j'ai été responsable avait parmi ses missions l'association d'attributs sémantiques à des objets mathématiques (cylindres par exemple) pour les rendre objets du monde mécanique ou *feature* (un trou par exemple). Un autre objectif était de définir et gérer des contraintes de paramétrage entre ces *features*, donnant à l'utilisateur la possibilité de modéliser numériquement des règles de conception mécanique : le positionnement d'un trou dont la circonférence doit être éloignée d'au moins une certaine distance du bord de la pièce ou dont le diamètre est contraint par l'outil de perçage. Un changement de spécification se traduisant par un changement des règles ou des géométries sous-jacentes, la pièce devait être reconstruite automatiquement en respectant les nouvelles spécifications. Cette automatisation était conçue pour éviter des modifications manuelles fastidieuses et garantir l'intégrité de la conception. L'intelligence artificielle étant en plein essor, c'est un moteur à base de règles sur des contraintes géométriques et dimensionnelles qui a été développé, avec ses avantages (règles déterministes et compréhensibles) et ses inconvénients (difficultés à définir les poids des différentes règles, difficultés de maintenance et de garantie de non régression d'une version à l'autre). Sous ma direction, le modèle SATT (Surfaces Associées Technologiquement et Topologiquement, ou *TTRS*) a été intégré. Il était directement issu des travaux de recherche en cours d'André Clément [2] (Desrochers et Clément, 1994; Clément et al., 1998), permettant de définir les informations de tolérancement dès le processus de conception.

Ce contact avec le monde de la recherche, l'envie d'avoir du temps pour développer des projets plus mathématiques, l'attrait pour le monde de l'aléatoire que je n'avais fait qu'effleurer pendant mes études m'ont amenée à entreprendre un projet personnel jusque-là laissé de côté : me lancer pleinement dans la recherche. Après un DEA en modélisation aléatoire, j'ai commencé à trente-et-un ans une thèse à mi-temps sous la direction d'Elisabeth Gassiat, soutenue quatre ans plus tard sous le titre *Test de Modèles par Maximum de Vraisemblance* (Keribin, 1999). Cette thèse a enraciné la thématique de mes travaux.

1. https ://www.3ds.com/fr/
2. Laboratoire LISMMA

J'y ai étudié le test de rapport de vraisemblance de l'ordre d'un modèle dans différentes configurations :

- observations i.i.d. d'un modèle identifiable même quand l'ordre est surestimé (Keribin et Haughton, 2003) ;
- observations i.i.d. d'un modèle de mélange (non identifiable quand le nombre de composantes est surestimé) ; une des conséquences étant la consistance du critère BIC pour ces modèles sous certaines conditions (Keribin, 1998, 2000)
- modèle de chaîne de Markov cachée à observations continues (Gassiat et Keribin, 2000)

J'ai alors porté la voix de la CAO au sein du Master Ingénierie Mathématique d'Orsay pendant trois années de contrat PAST à l'université Paris-Sud, mi-temps salariée à Dassault Systèmes et mi-temps enseignant chercheur à l'université. Sans renier mon histoire, cette période transitoire m'a convaincue de poursuivre mon projet universitaire, et j'ai été recrutée à l'IUT d'Orsay, puis au département mathématique de la faculté des sciences de l'université Paris Sud.

L'équipe INRIA-SELECT ayant étant créée au même moment, ma rencontre déterminante avec Gilles Celeux a marqué le début d'une longue et fructueuse collaboration : elle a mis au cœur de mes travaux la classification non supervisée (*clustering* puis à plus long terme *co-clustering*) et le choix de modèle, principalement sous l'angle des méthodes de modélisation probabiliste, et en particulier celle de l'utilisation des modèles de mélange abordés pendant ma thèse. Les méthodes de classification non supervisées sont une pierre angulaire des méthodes automatiques d'apprentissage. Elles possèdent des atouts importants pour traiter le développement incessant de la collecte et du stockage des données, en partie grâce à leur parcimonie permettant de les utiliser comme un outil de réduction de dimension. Elles sont de ce fait au cœur de l'actualité de l'intelligence artificielle et du traitement des données massives.

Un parcours de recherche n'est pas forcément linéaire, il se nourrit des interactions avec son environnement et de ses explorations. Ainsi, en a-t-il été pour le mien. Le plan de ce document n'est donc pas chronologique, mais il a pour but de tracer un parcours recomposé du *clustering* au *co-clustering*, essentiellement mais pas uniquement ancré par la modélisation par modèle de mélange. Il s'est enrichi d'incursions dans des domaines plus éloignés, souvent motivés par l'étude de données initiées par des collaborations : données génomiques, météorologiques, phylogénétiques, financières, ou de façon très récente, ferroviaires.

La discipline statistique, peut-être plus qu'aucune autre discipline mathématique même appliquée, permet de développer une gamme de compétences polyvalentes : compétences théoriques pour asseoir les concepts et valider leurs domaines d'utilisation ; méthodologiques pour concevoir des algorithmes efficaces, et proposer des modèles adaptés ; appliquées pour répondre aux attentes du monde actuel plus que jamais pourvoyeur de données à analyser. Mes travaux sont fidèles à cette vision : de nature théorique (identifiabilité de modèles, consistance de procédures,...), méthodologique (méthodes bayésiennes variationnelles, classification croisée, algorithmes d'estimation, ...) et appliquée (neuro-imagerie, pharmacovigilance).

L'optique d'écriture de ce mémoire n'est pas celle d'une revue exhaustive de littérature sur les sujets abordés, des articles très bien documentés existent et seront mentionnés, mais plutôt celle de mettre en contexte les recherches effectuées ou à venir.

Le chapitre 1 présente le contexte de la classification non supervisée, balaye rapidement différentes approches avant de s'attacher à celle des modèles de mélange fini. Il en rappelle les points saillants (identifiabilité, estimation, algorithme) et aborde la question fondamentale du choix du nombre de composantes. Un résultat important de mes travaux de thèse (Keribin, 1999, 2000) est rappelé. Une application récente en cancérologie (Keribin et al., 2019), issue d'une collaboration avec Yves Rozenholc (Université Paris Descartes) et Tatiana Popova (Institut Curie) propose un modèle original de modèle de mélange contraint sur des données de *micro-arrays* et discute l'utilisation pratique de méthodes de sélection de modèles (Keribin, 2019). La dernière section ouvre à l'utilisation du clustering de variables.

Le chapitre 2 étend le clustering à la classification croisée non supervisée (co-clustering) d'individus et de variables. Il présente le modèle des blocs latents, un modèle probabiliste généralisant le modèle de mélange i.i.d. classique. La structure de dépendance qu'il induit sur les données empêche le calcul numérique de la vraisemblance. D'importantes contributions théoriques (identifiabilité (Keribin et al., 2015), consistance des estimateurs (Mariadassou et al., 2016; Brault et al., 2019)) et méthodologiques (stratégies et algorithmes d'estimation, sélection de modèle (Keribin et al., 2010, 2012, 2015)) sont présentées et illustrées sur un exemple d'application. Une grande partie de ces travaux est directement ou indirectement issue de la thèse de Vincent Brault (Brault, 2014), que j'ai co-dirigée avec Gilles Celeux.

Le chapitre 3 part du besoin applicatif de la détection d'effets secondaires médicamenteux. Il développe l'aspect méthodologique en discutant l'utilisation du modèle des blocs latents Poissonien sur des données de contingence médicaments/effets, puis en définissant le modèle des blocs latents multiples pour prendre en compte les données individuelles (Robert et al., 2015). Ces données individuelles sont très clairsemées mais leur utilisation pourrait pallier les situations de co-prescription ou de masquage auxquels sont sujets les tableaux de contingence. Un algorithme glouton de parcours de modèle est mis en place (Robert et al., 2016). Ces travaux ont été menés dans le cadre de la thèse de Valérie Robert (Robert, 2017), co-dirigée avec Gilles Celeux et en collaboration avec Pascale Tubert-Bitter (INSERM). Ils illustrent l'utilisation du co-clustering comme un outil intéressant et utile de réduction de dimension (Keribin et al., 2017)

Le chapitre 4 retrace des travaux plus anciens portant sur une application de neuroimagerie par IRM fonctionnelle, et réalisés pendant la thèse de Vincent Michel (Michel, 2010) co-dirigée avec Gilles Celeux et Bertrand Thiron (Inria-Parietal). Le décodage du cerveau à partir de cartes d'activation est un problème d'apprentissage supervisé, mais la grande dimension des images (100 à 200 images de 10^5 voxels) nécessite de mettre en place des techniques régularisées, ou de réduire drastiquement la dimension tout en gardant des résultats interprétables. Après la présentation du cadre applicatif de l'IRM fonctionnelle, les limites des modélisations supervisées classiques sont discutées. Utilisant un a priori bayésien sur les poids des différentes variables, le modèle MSBR (*Multi Sparse Bayésian Regression*, Michel et al. (2010b, 2009, 2011)) entre dans la catégorie des méthodes régularisées. La méthode *Supervised clustering* (Michel et al., 2010a, 2012), mariant une méthode de clustering non supervisé contraint spatialement et l'utilisation d'un score de prédiction pour choisir une partition parmi celles construites, entre dans celle de la réduction de dimension.

Le chapitre 5 rassemble des projets qui partent tous de considérations applicatives, mais ne relèvent plus du champ du clustering. Trois sujets seront présentés, qui, à défaut d'être ou d'avoir abouti, illustrent comment des besoins applicatifs font surgir des questions théoriques ou méthodologiques intéressantes. Il s'agit (1) de la définition d'un modèle d'évolution d'arbres de phylogénie et la construction de tests, en collaboration avec Marie-Anne Poursat (Université Paris Sud), ayant donné lieu au développement d'un logiciel pour estimer le modèle *covarion* ; (2) de la prise en compte de variables météorologiques dans la prédiction de l'intensité de feux de forêts (Hernandez et al., 2015) en collaboration avec Charles Hernandez (Laboratoire de météorologie de l'Ecole Polytechnique) ; (3) de l'étude de l'agrégation d'experts sur des séries financières en collaboration avec Jean-Marc Guillard (Dassault-Systèmes et Stats4Trade).

Le chapitre 6 conclut le mémoire en ouvrant des pistes de recherche.

Chapitre 1

Clustering et modèle de mélange

Peut-on définir des groupes dans une population, et si oui, combien ? Ces questions font l'objet de la *classification non supervisée*. Elles sont récurrentes dans de nombreuses applications, comme par exemple la délimitation d'espèces biologiques, la segmentation d'image, la reconnaissance d'objet ou de parole ou la segmentation de clientèle.

Si on peut visuellement proposer de grouper des individus lorsque ceux-ci sont observés sur moins de trois variables, ces regroupements restent empiriques. De plus, l'accroissement incessant des données collectées rend nécessaire la définition de modèles et méthodes pour automatiser ce traitement.

Je commencerai par rappeler le cadre de la classification non supervisée, puis je présenterai le modèle de mélange comme un outil de modélisation probabiliste pour la classification non supervisée. Un exemple illustrera l'utilisation d'un modèle de mélange contraint pour définir la carte d'identité génétique d'une tumeur cancéreuse.

1.1 Cadre non supervisé

Les méthodes de *classification automatique* ou *clustering* permettent de partitionner ex nihilo des individus en groupes de façon automatique, et sans variable réponse identifiée dans l'échantillon. Ces méthodes d'apprentissage sont *non supervisées*, par opposition aux méthodes supervisées. Rappelons plus précisément la différence entre ces deux approches.

Dans un modèle d'apprentissage statistique supervisé, la variable réponse $y \in \mathcal{Y}$ doit être expliquée ou prédite en fonction de variables explicatives $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}$. Les variables de $\mathcal{X}$ sont qualitatives ou quantitatives. $\mathcal{Y}$ quantitatif ($\mathbb{R}$ par exemple) fait référence à un problème de régression, tandis que $\mathcal{Y}$ qualitatif fait référence à un problème de *classification supervisée* ou *méthode de discrimination*. À partir d'un échantillon d'apprentissage $\mathcal{D}_n = ((x_1, y_1), \ldots, (x_n, y_n))$, d'un prédicteur $f : \mathcal{X} \to \mathcal{Y}$, d'une fonction de coût mesurant la ressemblance entre $f(X)$ et Y (perte 0-1 par exemple : $\ell(f(X, Y)) = \mathbb{1}_{Y \neq f(X)}$) et d'un risque $\mathcal{R}(f) = \mathbb{E}(\ell(f(X, Y))$, l'apprentissage supervisé permet d'apprendre une règle $\widehat{f}$ à partir de l'échantillon d'apprentissage $\mathcal{D}_n$, telle que le risque $\mathcal{R}(\widehat{f})$ soit petit en moyenne, ou avec grande probabilité sur $\mathcal{D}_n$. Le problème de classification supervisé est bien posé et on y retrouve les ingrédients identifiés par Tom Mitchell, Carnegie Mellon University [1].

En classification non supervisée, les observations ne possèdent pas de variable à expliquer, et $\mathcal{D}_n = (x_1, \ldots, x_n)$. La fonction f est définie de $\mathcal{X}$ dans $\mathcal{Z} = \{1, \ldots, K\}$ où K est

1. *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*

un nombre fixé de clusters. Le résultat du clustering est un vecteur $\mathbf{z} = (z_1, \ldots, z_n) \in \mathcal{Z}$ indiquant l'allocation de chaque observation à un cluster. Ici, seul $\mathcal{D}_n$ est bien identifié, ce qui rend le problème par essence mal posé. Le résultat d'un clustering dépend en particulier du choix du critère de ressemblance (ou de dissemblance) entre deux individus et entre deux groupes d'individus. Plus généralement, l'utilisateur doit faire face à un grand nombre d'options qui ont une influence sur la validation des résultats et leur l'interprétation (Milligan, 1996) : choix des objets/individus, des variables, de leur transformation, choix d'une mesure de (dis)-similarité, choix d'une méthode de clustering, choix/détermination du nombre de clusters. Nous évoquons ces deux derniers points dans les paragraphes suivants.

Remarque Le terme classification, si on ne le complète pas, est ambigu en français, car il fait référence aussi bien à la classification supervisée que non supervisée. Le terme anglais lève toute ambiguïté : classification (dans le cas supervisé) et clustering (dans le cas non supervisé). C'est pour lever toute confusion que nous préfèrerons utiliser cet anglicisme. Dans la suite, classe, cluster et groupe sont synonymes.

1.1.1 Méthodes classiques

Il existe de nombreuses méthodes de clustering, dont le choix dépend de l'objectif poursuivi. L'approche *réaliste* permet d'éclairer la façon dont la structure sous-jacente ("vraie structure") est réellement connectée aux données. De fait, il s'agit bien d'une hypothèse de modélisation, discutée en amont en fonction de certaines connaissances a priori, et qui découle de l'idée que l'on se fait d'un cluster du phénomène observé. L'approche *constructive* permet quant à elle de partitionner les données en groupes sans s'intéresser forcément à la réalité des groupes formés, qui sont simplement définis comme une étape, en général de simplification, d'un processus ultérieur. Dans ce cas, les caractéristiques du cluster sont définies sans faire référence à une "vraie" structure. La démarcation entre les deux types d'approches n'est cependant pas nette, et Hennig et al. (2015) indique par exemple que le clustering social peut être vu comme la détermination de classes qui préexistent (vision marxiste) ou comme un outil commode pour structurer les différences.

La définition des clusters peut prendre en compte des caractéristiques telles que les dissimilarités intra-cluster, inter-cluster, l'homogénéité à un modèle probabiliste, la forme de l'espace contenant un cluster ou la taille des clusters (cf. Hennig et al. (2015) pour une liste plus étendue). Le choix de la méthode dépend de son aptitude à s'accorder aux critères retenus.

Les méthodes se distinguent également par le type d'outil mathématique utilisé : déterministes (méthodes de distance comme K-means ou K-medoids, classification hiérarchique, décomposition spectrale), ou basées sur des modèles probabilistes (modèles de mélange, méthodes non paramétriques), voir Friedman et al. (2001) pour une introduction à ces méthodes.

1.1.2 Combien de groupes ?

La classification obtenue dépend évidement du nombre de groupes. Le cadre non supervisé étant mal posé, le problème pour décider du nombre de clusters est lui-même délicat, et peut ne pas avoir de réponse unique. Avec une méthode déterministe, un seuil a priori (et donc arbitraire) du critère est utilisé. L'avantage d'utiliser un modèle probabiliste, et en particulier un modèle de mélange, est de bénéficier du cadre de la théorie de la sélection

de modèle. Ainsi, le critère asymptotique de choix de modèle BIC (Bayesian Information Criteria, Schwarz et al. (1978)), est consistant dans le cas de certaines familles de mélange quand le vrai modèle appartient à la collection de modèle (Keribin, 2000), mais peut surestimer le nombre de composantes si ce n'est pas le cas. C'est un critère d'identification. Le critère ICL (*Integrated Completed Likelihood*, Biernacki et al. (2000)) permet de tenir compte de la difficulté de la situation de mélange. Le critère n'est pas consistant, mais développe une vision classification. Nous reviendrons sur ces critères dans la section 1.3

1.1.3 Performance des méthodes

Comment peut-on comparer des méthodes de clustering ? Avec le manque d'information a priori, il est difficile de définir les performances et de comparer des méthodes qui n'ont souvent pas les mêmes objectifs.

Dans une approche d'identification, des méthodes de simulation permettent de confronter une partition $\mathbf{z} = (z_1, \ldots, z_n)$ obtenue par une méthode avec la vraie partition $\mathbf{z}^\star = (z_1^\star, \ldots, z_n^\star)$ connue (puisque simulée). La performance du clustering peut alors être établie en utilisant l'erreur de classification

$$e_{\mathbf{z},\mathbf{z}^\star} = \max_{\mathbf{z}' \sim \mathbf{z}} \sum_{i=1}^{n} \mathbb{I}_{z_i' \neq z_i^\star} \tag{1.1}$$

Ce critère est dépendant de la numérotation des clusters eux-mêmes. Pour l'en rendre indépendant, le risque estimé de la méthode est défini par l'erreur minimum sur l'ensemble des permutations. Ce critère semble naturel ici, mais peut se heurter à un problème combinatoire quand le nombre de classes est important.

L'*Ajusted Random Index* (ARI) s'affranchit naturellement de la numérotation des classes en comparant les paires d'observations (Hubert et Arabie, 1985; Youness et Saporta, 2004) D'autres critères peuvent être utilisés, résumant en particulier la qualité des clusters obtenus (homogénéité intra classe, séparation des classes par exemple).

1.2 Modèles de mélange paramétrique fini

Les modèles de mélange posent un modèle *probabiliste* pour résoudre un problème de clustering et donnent lieu une abondante littérature (McLachlan et Basford (1988); Lindsay (1995); Biernacki (2017) entre autres). Leur flexibilité en font un outil de choix.

Soit $\mathcal{F} = (\varphi(.;\alpha))_{\alpha \in \Xi}$ une famille de densités paramétriques par rapport à une mesure ν sur $\mathcal{X}$, de paramètre $\alpha \in \Xi$. La loi d'une variable aléatoire sur $\mathcal{X}$ suit un modèle de mélange de lois de $\mathcal{F}$ si et seulement si sa densité par rapport à ν s'écrit

$$f(x) = \int \varphi(x;\alpha) d\xi(\alpha) \;\; d\nu - \text{ presque sûrement}$$

où ξ est une loi de probabilité sur l'espace du paramètre, appelée distribution du mélange. Le mélange est dit *fini* quand la distribution de mélange est discrète, ne chargeant que K valeurs : $\xi = \sum_{k=1}^{K} \pi_k \delta_{\alpha_k}$ où $\{\alpha_1, \ldots, \alpha_K\}$ est le support de ξ, δ_α la mesure de Dirac en α et $\pi_k \in [0;1]$ tels que $\sum_{k=1}^{K} = 1$:

$$f(x) = \sum_{k=1}^{K} \pi_k \varphi(x;\alpha_k) \tag{1.2}$$

Les densités $\varphi(.;\alpha_k)$ sont appelées *composantes* du mélange et les $(\pi_1,\ldots,\pi_K)=\pi$ sont les *proportions* ou *poids* du mélange. On appelle Π_K l'ensemble des K-uplets de poids. Un modèle de mélange de lois $\mathcal{F}$ est défini pour K fixé par

$$\mathcal{M}_K=\left\{\sum_{k=1}^{K}\pi_k\varphi(x;\alpha_k)|(\pi,\alpha_1,\ldots,\alpha_K)\in\Theta_K\right\} \tag{1.3}$$

avec $\Theta_K\subset\Pi_K\times\Xi^K$.

Dans un modèle statistique de mélange fini, les observations d'un échantillon $X=(X_1,\ldots,X_n)$ de taille n sont indépendantes et issues de la même loi de mélange f dont tout ou partie des paramètres sont inconnus et à inférer. Quand $\mathcal{F}$ est la famille des lois gaussiennes par exemple, le mélange est dit gaussien, les paramètres des gaussiennes et les poids du mélange sont à identifier. Un échantillon d'un modèle gaussien à trois composantes est représenté sur la figure 1.1.

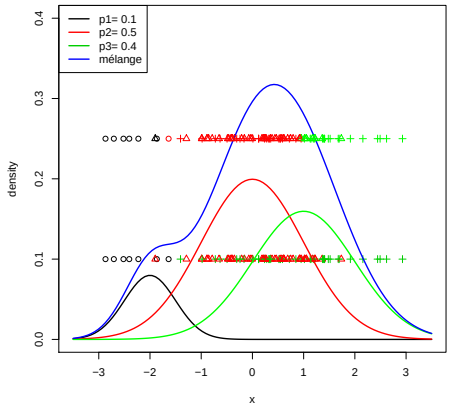


FIGURE 1.1 – Loi de mélange de trois gaussiennes univariées ; les observations sont représentées sur la ligne 0.1, de forme et couleur dépendant de la composante dont elles sont issues. Sur la ligne 0.2, le clustering est estimé.

1.2.1 Identifiabilité

On rappelle que la famille paramétrique $\mathcal{F}=\{\varphi(x;\alpha),\alpha\in\xi\}$ indexée par $\alpha\in\Xi$ est identifiable si et seulement si φ vue comme fonction de α est injective. Or, dans le cas des mélanges, il suffit d'intervertir l'ordre des classes pour mettre en défaut la définition : même si la classe $\mathcal{F}$ est identifiable, le mélange ne l'est pas à cause de $K!$ permutations possibles. La définition d'identifiabilité est alors affaiblie : un modèle de mélange est (faiblement) identifiable si le paramètre est la loi de probabilité sur Ξ du mélange discret (Teicher et al., 1960) :

$$\sum_{k=1}^{K}\pi_k\varphi(x;\alpha_k)=\sum_{k=1}^{K}\pi'_k\varphi(x;\alpha'_k)\ \nu p.s.\ \text{ssi}\ \sum_{k=1}^{K}\pi_k\delta_{\alpha_k}=\sum_{k=1}^{K}\pi'_k\delta_{\alpha'_k},$$

la deuxième égalité étant prise au sens des mesures de probabilité sur Ξ.

La plupart des modèles de mélanges finis de lois continues sont identifiables sous cette définition quand les poids sont strictement positifs et les paramètres des composantes différents, sauf les mélanges de la loi uniforme. Ce n'est pas le cas pour les lois discrètes :

le mélange de lois binomiales ne sont pas identifiables si le nombre d'observations est trop faible par rapport au nombre de variables, ie $n < 2K - 1$ (Gyllenberg et al., 1994).

Un autre problème d'identifiabilité survient quand le nombre de composantes est surestimé. En effet dans cas, pour $\varepsilon \in [0; \pi]$,

$$\pi\varphi(x;\alpha_1) + (1-\pi)\varphi(x;\alpha_2) = \varepsilon\varphi(x;\alpha_1) + (\pi-\varepsilon)\varphi(x;\alpha_1) + (1-\pi)\varphi(x;\alpha_2).$$

La paramétrisation localement conique de Dacunha-Castelle et Gassiat (1997) a permis de prendre en compte ce problème, et nous l'utiliserons pour définir des critères de choix, voir section 1.3.

1.2.2 Estimation

Dans cette section, on suppose que K est fixé et que la "vraie" distribution se trouve dans la famille $\mathcal{M}_K$. Les paramètres à estimer sont $\theta = (\pi, \alpha_1, \ldots, \alpha_K)$, éventuellement $\theta = (\pi, \eta)$ si les paramètres des densités $\alpha_k = \alpha_k(\eta)$ s'expriment eux-mêmes par rapport à un paramètre η. C'est le cas par exemple d'un mélange gaussien où la variance est commune à l'ensemble des composantes.

La loi des observations étant définie, la méthode naturelle d'estimation est celle par maximum de vraisemblance :

$$\widehat{\theta} \in \arg\max_{\theta} \prod_i f(x_i;\theta). \tag{1.4}$$

Cette estimation pose des problèmes, le maximum pouvant ne pas exister, ni la vraisemblance être bornée (par exemple, quand la variance d'une composante gaussienne n'est pas bornée inférieurement par un nombre strictement positif). Restreindre la définition de l'espace des paramètres pour garantir l'existence (Redner, Hathaway, Lindsay) peut permettre de s'en affranchir.

D'un point de vue pratique, l'estimation par maximum de vraisemblance utilise l'algorithme *Expectation-Maximisation* (EM, Dempster et al. (1977)). Cet algorithme itératif permet de trouver l'estimateur du maximum de vraisemblance d'un modèle contenant des variables *latentes* non observées. Ainsi est introduite pour chaque observation X_i la variable latente Z_i représentant l'indice (ou *label*, ou assignation) de la composante d'appartenance de cette observation : $Z_i \sim \mathcal{M}(1,\pi)$ et la densité de $X_i|\{Z_i = k\}$ est $\varphi(.,\alpha_k)$. Les labels Z_i sont indépendants.

Le principe de l'algorithme EM repose sur le fait de considérer la vraisemblance complète (variables observées et latentes) :

$$p(\mathbf{x},\mathbf{z};\theta) = \prod_i \mathbb{P}(Z_i = k)\varphi(x_i;\alpha_k) = \prod_i \prod_k [\pi_k\varphi(x_i;\alpha_k)]^{z_{ik}}$$

où z_{ik} est une variable binaire codant l'appartenance de l'observation i à la classe k. Ainsi

$$\log p(\mathbf{x},\mathbf{z};\theta) = \sum_i \sum_k z_{ik} \left[\log(\pi_k\varphi(x_i;\alpha_k))\right]$$

La loi du label Z_i conditionnellement aux observations est une loi multinomiale $\mathcal{M}(t_{i1},\ldots,t_{iK})$ définie par :

$$t_{ik}(\theta) := \mathbb{E}(z_{ik}|\mathbf{x};\theta) = \mathbb{P}(Z_i = k|x_i;\theta) = \frac{\pi_k\varphi(x_i;\alpha_k))}{\sum_\ell \pi_\ell\varphi(x_i;\alpha_\ell)} \tag{1.5}$$

L'algorithme EM résout l'optimisation en itérant deux étapes

Algorithme 1 (EM)**.** *Après initialisation, itérer jusqu'à convergence*

— *Espérance : prend en compte l'addition des données manquantes en calculant l'espérance de la log-vraisemblance des données complètes (observées et latentes) conditionnellement à la loi des observations sous le paramètre en cours* $\theta^{(c)}$*,*

$$Q(\theta|\theta^{(c)}) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{z}; \theta)|\mathbf{x}; \theta^{(c)}] = \sum_i \sum_k t_{ik}^{(c)} \left(\log(\pi_k) + \log(\varphi(x_i; \alpha_k)\right)$$

— *Maximisation : maximise l'expression précédente en* θ*,*

$$\theta^{(c+1)} = \arg\max_{\theta} Q(\theta|\theta^{(c)})$$

d'où, en particulier,

$$\pi_k^{(c+1)} = \frac{\sum_i t_{ik}^{(c)}}{n}.$$

L'estimateur du maximum de vraisemblance est la valeur θ^{fin} *obtenu à la dernière itération.*

Cet algorithme a la propriété d'augmenter la vraisemblance à chaque itération. On a :

$$\begin{aligned}
\log p(\mathbf{x}; \theta) &= \log p(\mathbf{x}, \mathbf{z}; \theta) - \log p(\mathbf{z}|\mathbf{x}; \theta) \\
&= \mathbb{E}[\log p(\mathbf{x}, \mathbf{z}; \theta)|\mathbf{x}; \theta^{(c)}] - \mathbb{E}[\log p(\mathbf{z}|\mathbf{x}; \theta)|\mathbf{x}; \theta^{(c)}] \\
&= Q(\theta|\theta^{(c)}) - H(\theta|\theta^{(c)})
\end{aligned}$$

Soit $\tilde{\theta} \in \arg\max_\theta Q(\theta|\theta^{(c)})$, alors $\tilde{\theta}$ fait augmenter la vraisemblance :

$$L(\tilde{\theta}) - L(\theta^{(c)}) = Q(\tilde{\theta}|\theta^{(c)}) - Q(\theta|\theta^{(c)}) + H(\theta^{(c)}|\theta^{(c)}) - H(\tilde{\theta}|\theta^{(c)}) \geq 0$$

car $H(\theta^{(c)}|\theta^{(c)}) - H(\tilde{\theta}|\theta^{(c)}) \geq 0$ est la divergence de Kullback entre les deux lois.

La vraisemblance d'un modèle de mélange est multimodale et l'algorithme EM peut ne converger que vers un maximum local. Il est donc en général nécessaire de le relancer plusieurs fois avec différentes initialisations, puis de prendre la solution de plus grande vraisemblance. Des stratégies élaborées d'initialisation ont été notamment proposées par Biernacki et al. (2003).

Clustering et variante CEM Les $t_{ik}(\widehat{\theta})$ calculés à la dernière itération de l'algorithme EM estiment la loi des labels conditionnellement aux observations. Il est donc naturel d'affecter après convergence, chaque observation i à une composante par la règle du Maximum A Posteriori (*MAP*)

$$\widehat{z}_i = \arg\max_{k=1,\dots,K} t_{ik} \tag{1.6}$$

La figure 1.1 représente la classification ainsi estimée pour un modèle de mélange gaussien univarié.

L'algorithme EM, maximisant la vraisemblance, adresse l'aspect identification de la loi de mélange, et l'utilisation de la règle du MAP est une façon d'en dériver une partition. Quand l'objectif poursuivi est celui de la classification, il est pertinent de vouloir maximiser la vraisemblance complète $p(\mathbf{x}, \mathbf{z}, \theta)$ en θ et $\mathbf{z}$ qui prend le statut de paramètre inconnu

plutôt que celui de variable latente. Dans ce cas, s'ajoute à l'étape E celle d'une étape de classification (C) utilisant la règle du MAP (Celeux et Govaert, 1992) :

$$z_i^{(c)} = \arg \max_{k=1,\dots,K} t_{ik}^{(c)}.$$

L'étape M maximise à son tour $p(\mathbf{x}, \mathbf{z}^{(c)}, \theta)$ en θ. Cet algorithme a l'avantage de converger en un nombre fini d'itérations, mais il est sensible à l'initialisation. Quand la loi des composantes est gaussienne de variance commune sphérique, l'algorithme CEM est équivalent à celui des K-moyennes.

SEM : une variante stochastique L'algorithme EM est non seulement sensible à l'initialisation, mais il peut également rencontrer des cas de convergence extrêmement lente, ce qui peut rendre son utilisation difficile. Celeux et Diebolt (1986) ont proposé un algorithme EM augmenté d'une étape probabiliste (S) : lors de l'étape E, en chaque observation i, est tirée une multinomiale

$$z_i^{(c)} \sim \mathcal{M}(1, (t_{ik}^{(c)})_{k=1,\dots,K}),$$

générant une partition $\mathbf{z}^{(c)}$ des observations. L'étape M maximise la vraisemblance complète $p(\mathbf{x}, \mathbf{z}^{(c)}; \theta)$. Cet algorithme ne converge pas vers une valeur du paramètre : le type de convergence obtenu est une convergence en loi correspondant à la stationnarité de la suite des estimés autour d'une valeur moyenne. Cet algorithme converge notablement plus vite que l'algorithme EM et les perturbations introduites à chaque itération par les tirages aléatoires empêchent la convergence vers un maximum local instable de la vraisemblance. SEM peut aussi être vu comme une adaptation stochastique de l'algorithme CEM.

1.2.3 Interprétation variationnelle de l'EM

Nous revenons dans cette section sur une interprétation connue de l'EM et qui sera utilisée dans la suite de ce mémoire. Soit q une loi quelconque appartenant à l'ensemble des lois définies sur les labels Z_i. On peut écrire à partir de la vraisemblance

$$\begin{aligned}
\log p(\mathbf{x}; \theta) &= \log(p(\mathbf{x}, \mathbf{z}; \theta)/q(\mathbf{z})) - \log(p(\mathbf{z}|\mathbf{x}; \theta)/q(\mathbf{z})) \\
&= \mathbb{E}\left[\log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} | \mathbf{x}; \theta\right] - \mathbb{E}\left[\log \frac{p(\mathbf{z}|\mathbf{x}; \theta)}{q(\mathbf{z})} | \mathbf{x}; \theta\right] \\
&= \mathcal{F}(q, \theta) + KL(q, p(.|\mathbf{x}; \theta))
\end{aligned}$$

où $KL(q, p(.|\mathbf{x}; \theta))$ est la divergence de Kullback entre la loi q et la loi conditionnelle des labels $p(.|\mathbf{x}; \theta)$. Le terme $\mathcal{F}(q, \theta)$ est généralement appelé *énergie libre* et q *loi libre*. Considérant toutes les lois q possibles,

$$p(.|\mathbf{x}; \theta) = \arg \min_q KL(q, p(.|\mathbf{x}; \theta)) = \arg \max_q \mathcal{F}(q, \theta) \tag{1.7}$$

À l'optimum, l'énergie libre est nulle et la divergence de Kullback est égale à la vraisemblance observée. Ainsi, l'étape E de l'algorithme EM peut être vue comme une étape d'optimisation variationnelle : le calcul de l'espérance est remplacé par une optimisation. L'algorithme EM est une optimisation alternée de l'énergie libre en q et θ : c'est l'*interprétation variationnelle* de l'algoritme EM.

Cela ne change rien pour l'estimation des mélanges simples, mais cette interprétation facilite la définition d'algorithmes alternatifs quand l'espérance de l'étape E n'est pas directement calculable. En restreignant l'optimisation en q à un espace Q de lois pour lesquelles le calcul de l'espérance est réalisable :

$$p(.|\mathbf{x};\theta) \simeq q^*(.|x) = \arg\min_{q\in\mathcal{Q}} KL(q,p) = \arg\max_{q\in\mathcal{Q}} \mathcal{F}(q,\theta), \tag{1.8}$$

le résultat de l'étape E variationnelle devient une fonction approchée de $p(.|x;\theta)$: c'est le principe de l'*approximation variationnelle*. À l'optimum q^*, la divergence de Kullback est non nulle et l'énergie libre un minorant de la log-vraisemblance : $\mathcal{F}(q,\theta) < \log p(x;\theta)$. Le calcul de l'emv est approché : c'est l'estimateur variationnel

$$\widehat{\theta}^{VAR} \in \arg\max_{\theta} \max_{q\in\mathcal{Q}} \mathcal{F}(q,\theta). \tag{1.9}$$

1.2.4 Algorithme bayésien variationnel

Le paradigme bayésien considère le paramètre θ comme étant lui-même aléatoire et de loi est à inférer. A partir d'une loi a priori $p(\theta)$ et d'une vraisemblance $p(\mathbf{x}|\theta)$, le théorème de Bayes permet d'obtenir la loi a posteriori $p(\theta|\mathbf{x})$, tirant partie de la connaissance apportée par la vraisemblance observée. En présence de données manquantes, des lois a priori sont définies sur les paramètres $p(\theta)$ et les labels $p(\mathbf{z})$. L'EM bayésien variationnel permet l'obtention de la loi jointe a posteriori $p(z,\theta|x)$. Quand celle-ci n'est pas calculable, une approximation variationnelle est faite en optimisant en général parmi les lois jointes factorisées :

$$p(\theta,\mathbf{z}|\mathbf{x}) \simeq q_\theta(\theta)q_z(\mathbf{z}) = q_\theta(\theta)\prod_i q_i(z_i).$$

Si les observations sont i.i.d., la loi $q_z(\mathbf{z})$ se factorise également sans approximation supplémentaire, aboutissant à l'algorithme Variational Bayes complet (Beal et al., 2003; Bishop, 2006) :

Algorithme 2 (Full-VBayes)**.** *Après initialisation, itérer jusqu'à convergence :*

— *Etape VBE : mettre à jour $q(\mathbf{z}|\mathbf{x})$, l'approximation variationnelle de la loi a posteriori des variables latentes :*

$$q_{\mathbf{z}}^{c+1} = \arg\max_{q_{\mathbf{z}}} \mathcal{F}(q_{\mathbf{z}}, q_\theta^c)$$

— *Etape VBM : mettre à jour $q(\theta|\mathbf{x})$, l'approximation variationnelle de loi a posteriori des paramètres*

$$q_\theta^{c+1} = \arg\max_{q_\theta} \mathcal{F}(q_{\mathbf{z}}^{c+1}, q_\theta) = \arg\max_{\theta} = Q(\theta|\theta^{c+1}).$$

Si l'objectif est seulement d'atteindre le *mode* de la loi a posteriori en θ,

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathbf{x}) \tag{1.10}$$

l'algorithme peut être adapté en notant que

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{x}).$$

Ainsi, seule l'étape M est modifiée, amenant à l'algorithme suivant :

Algorithme 3 (EM-VBayes)**.** *Après initialisation, itérer jusqu'à convergence :*
— *Etape E : mettre à jour* $p(\mathbf{z}|\mathbf{x};\theta)$ *:*

$$q_{\mathbf{z}}^{c+1} = \arg\max_{q_{\mathbf{z}}} \mathcal{F}(q_{\mathbf{z}}, \theta^c); \;\; Q(\theta|\theta^{c+1}) = \mathcal{F}(q_{\mathbf{z}}^{c+1}, \theta)$$

— *Etape M : maximiser*

$$\theta^{c+1} = \arg\max_{\theta} \mathcal{F}(q_{\mathbf{z}}^{c+1}, \theta) = \arg\max_{\theta} \left(Q(\theta|\theta^c) + \log p(\theta)\right)$$

L'estimateur $\widehat{\theta}_{MAP}$ *est la valeur* θ^{fin} *obtenue à la dernière itération.*

L'étape M force à augmenter $p(\theta|\mathbf{x})$ (McLachlan et Krishnan, 2007, chap. 6, p. 231). La prise en compte d'a priori bayésien permet de régulariser l'estimation, et en particulier de combattre le phénomène de dégénérescence de classes dont je parlerai à la section 2.5.2. Il est déterministe et converge comme l'EM vers un extremum local.

Il est difficile d'évaluer théoriquement la précision de ces méthodes, puisque la loi a posteriori exacte n'est en général pas calculable hors quelques cas simples. Or celles-ci sont utilisées dans de nombreuses applications. C'est la lecture d'articles les utilisant en neuroimagerie (chapitre 4) qui m'a amenée à écrire une synthèse bibliographique des résultats théoriques sur les méthodes bayésiennes variationnelles (Keribin, 2010). En quelques mots, les méthodes variationnelles peuvent être convergentes quand il n'y a pas trop de données latentes, avec une variance limite qui peut être inférieure à la variance réelle. Mais la consistance est perdue quand la proportion de variables latentes devient trop importante ou quand le bruit est trop important. De façon générale, il est plus facile d'approcher la localisation du mode que la valeur au mode : la valeur de la fonctionnelle au mode est souvent assez mal estimée, même si la localisation du mode est correct. Il s'agit donc d'être précautionneux lors de l'utilisation de l'énergie libre comme approximation de la vraisemblance dans les critères pénalisés de choix de modèle.

1.2.5 Échantillonnage de Gibbs

L'échantillonnage de Gibbs (Robert, 2007) est une alternative aux algorithmes variationnels. En particulier, quand les lois a priori sont conjuguées, il permet de simuler la loi a posteriori sans approximation. Sa difficulté réside dans la définition du temps de chauffe et de son critère d'arrêt. Il sera utilisé au chapitre 2.5.2.

1.3 Sélection du nombre de composantes

Jusqu'à présent le nombre de composantes K était supposé connu, ce qui n'est généralement pas le cas dans les applications. Nous balayons ici quelques méthodes de choix de modèle pour le nombre de composantes d'un modèle de mélange fini.

1.3.1 Vraisemblance pénalisée et critère BIC

Pendant ma thèse, j'ai étudié l'estimation de l'ordre (c'est-à-dire le nombre de composantes distinctes q^*) d'un modèle de mélange fini, en m'intéressant plus particulièrement aux critères asymptotiques de maximum de vraisemblance pénalisée. Pour tout ordre $q \in \{1, .., Q\}$, soit $\mathcal{L}(\theta_q) = \sum_{i=1}^n \log f(x_i; \theta_q)$ la log-vraisemblance d'un n-échantillon d'un modèle de mélange à q composantes. La statistique du maximum de vraisemblance

$$\mathcal{L}(\widehat{\theta}_q) = \sup_{\theta_q \in \Theta_q} l_n(\theta_q)$$

augmente avec l'ordre et un critère de choix de modèles doit compenser cette augmentation pour éviter le surajustement. Soit $a_{n,q} \in \mathbb{R}^+$ une séquence de réels positifs. La log-vraisemblance pénalisée est définie par

$$W_n(\mathcal{M}_q) = \mathcal{L}(\widehat{\theta}_q) - a_{n,q}.$$

L'idée d'utiliser un tel terme de pénalisation a été introduite pour la première fois par Akaike (1970) avec le critère AIC pour lequel la compensation est $a_{n,q} = dim(\mathcal{M}_q)$ où $dim(\mathcal{M}_q)$ est le nombre de paramètres libres estimés dans le modèle $\mathcal{M}_q$. Schwarz et al. (1978), suivant des considérations bayésiennes, a défini le critère BIC pour les familles exponentielles :

$$BIC(q) = \mathcal{L}(\widehat{\theta}_q) - \frac{dim(\mathcal{M}_q)}{2}\log(n). \tag{1.11}$$

Soit $p(\theta_q)$ une loi a priori sur le paramètre θ_q dans le modèle $\mathcal{M}_q$. Le critère BIC résulte de l'approximation asymptotique (par approximation de Laplace) de la vraisemblance intégrée :

$$p(\mathbf{x}; \mathcal{M}_q) = \int p(\theta_q, \mathbf{x}) d\theta_q = \int \mathcal{L}(\theta_q) p(\theta_q) d\theta_q.$$

Dans le cas des modèles de mélange, l'estimateur pénalisé $\hat{q}_n$ du (vrai) ordre q^* est

$$\hat{q}_n = \arg \max_{q \in \{1,\ldots,Q\}} (W_n(\mathcal{M}_q)).$$

Leroux (1992) a prouvé, sous certaines hypothèses du modèle et de la séquence $a_{n,q}$, que $\hat{q}_n$ ne sous-estime pas asymptotiquement le nombre de composantes presque sûrement. Nous avons étendu ces résultats (Keribin, 1998, 2000) :

Théorème 1 (Keribin (1998, 2000))**.** *Soit un modèle de mélange identifiable de vrai ordre q^* inconnu, satisfaisant les hypothèses de régularité de Dacunha-Castelle et Gassiat (1997).*

On suppose de plus qu'un majorant Q de l'ordre q^* est connu et que la séquence de pénalisation $a_{n,q}$ vérifie pour tout $q^* < q \leq Q$ les trois conditions suivantes :

$$a_{n,q+1} \geq a_{n,q} > 0, \quad a_{n,q} = o(n), \quad \lim_{n \to +\infty} a_{n,q}/a_{n,q^*} > 1$$

Alors,

- *l'estimateur $\hat{q}_n$ de l'ordre q^* est consistant*
- *$\hat{q}_n$ ne sous-estime pas q^* presque sûrement*
- *si de plus la séquence vérifie $\lim_{n \to +\infty} \log\log n / a_{n,q} = 0$, alors $\hat{q}_n$ et $\widehat{\theta}_{\hat{q}_n}$ sont presque sûrement consistants*

La preuve s'appuie sur la paramétrisation localement conique développée par Dacunha-Castelle et Gassiat (1997) pour montrer la convergence en loi du maximum de vraisemblance et identifier la loi limite. Les hypothèses du théorème sont en particulier vérifiées pour les mélanges finis gaussiens, dont les paramètres d'espérance appartiennent à un espace compact, les matrices de covariances sphériques et bornées inférieurement, montrant la consistance de BIC dans ce cas. Ce résultat a été étendu ultérieurement par Gassiat et Van Handel (2013) en supprimant l'hypothèse d'un majorant connu du nombre de groupes.

Cependant ces propriétés ne se maintiennent pas au cas des mélanges à régime markovien à cause de la double difficulté de la non identifiabilité et de la dépendance markovienne. Le résultat marquant est le comportement divergent de la statistique du rapport de vraisemblance dans le test d'un modèle i.i.d. contre un modèle de mélange de deux populations à régime markovien (Gassiat et Keribin, 2000).

1.3.2 Critère ICL

Les propriétés asymptotiques de BIC mentionnées précédemment ne sont valables que si le vrai modèle appartient à la collection de modèles. Sinon, BIC peut surestimer le nombre de composantes dans des situations mal spécifiées, comme illustré dans les simulations de Biernacki et al. (2000) pour lesquelles la vrai loi n'appartient à aucun des modèles considérés. Ceci peut s'expliquer par le fait que BIC sélectionne un modèle qui minimise la divergence de Kullback avec la vraie distribution. Dans ce sens, BIC est un critère d'*identification*.

Biernacki et al. (2000) ont proposé un critère permettant de s'affranchir des difficultés de BIC à sélectionner un modèle pertinent pour le choix du nombre de composantes. L'idée est de remplacer la vraisemblance observée par la vraisemblance complète $p(\mathbf{x}, \mathbf{z}; \mathcal{M})$, afin de prendre en compte la qualité du clustering généré et éviter les phénomènes de surestimation précédents :

$$p(\mathbf{x}, \mathbf{z}; \mathcal{M}) = \log \int p(\mathbf{x}, \mathbf{z}|\theta) p(\theta) d\theta \tag{1.12}$$

$$\simeq \max_{\theta} p(\mathbf{x}, \mathbf{z}|\theta) - \frac{dim(\mathcal{M})}{2} \log(n) + O(1). \tag{1.13}$$

Quand la loi a priori $p(\theta)$ est conjuguée, l'expression (1.12) est analytique et calculable. Sinon, elle est approchée asymptotiquement en utilisant une approximation de Laplace et devient (1.13). Comme les labels $\mathbf{z}$ ne sont pas observés, ces auteurs proposent de les affecter par une règle du MAP : $\widehat{\mathbf{z}}^{MAP}(\widehat{\theta}^{MLE})$. De plus, ils remplacent $\arg\max_\theta p(\mathbf{x}, \mathbf{z}|\theta)$ par l'estimateur du maximum de vraisemblance $\widehat{\theta}^{MLE}$. D'où l'expression du critère ICL

$$\begin{aligned} ICL(\mathcal{M}_q) &= p(\mathbf{x}, \widehat{\mathbf{z}}^{MAP}; \widehat{\theta}_q^{MLE}) - \frac{dim(\mathcal{M}_q)}{2} \log(n) \\ &= \mathcal{L}(\mathbf{x}; \widehat{\theta}_q^{MLE}) + \sum_{i=1}^{n} \sum_{k=1}^{q} \widehat{\mathbf{z}}_{ik}^{MAP} \log t_{ik}(\widehat{\theta}_q^{MLE}) - \frac{dim(\mathcal{M}_q)}{2} \log(n). \end{aligned} \tag{1.14}$$

McLachlan et Peel (2000)[6.10.3] proposent plutôt de remplacer $\mathbf{z}$ par $t_{ik}(\widehat{\theta}_q^{MLE})$, faisant apparaître un terme d'entropie mesurant la confiance dans la partition définie.

$$ICL_2(\mathcal{M}_q) = \mathcal{L}(\mathbf{x}; \widehat{\theta}_q^{MLE}) + \sum_{i=1}^{n} \sum_{k=1}^{q} t_{ik}(\widehat{\theta}_q^{MLE}) \log t_{ik}(\widehat{\theta}_q^{MLE}) - \frac{dim(\mathcal{M}_q)}{2} \log(n). \tag{1.15}$$

Ainsi, ICL n'est pas conçu pour sélectionner un modèle optimal au sens de la divergence de Kullback, mais un modèle menant à une classification pertinente.

Le critère ICL sera adapté au co-clustering dans les chapitres 2 et 3.

1.3.3 Heuristique de pente

De façon plus récente Birgé et Massart (2001), Birgé et Massart (2007) ont développé une approche non asymptotique alternative. Cette méthode pilotée par les données permet

de calibrer une constante multiplicative $\kappa > 0$ dans le critère pénalisé défini dans notre contexte comme suit :

$$crit(q;\kappa) = -\mathcal{L}(\widehat{\theta}_q;\mathbf{x}) + \ \kappa\, D_q$$

où D_q reflète la complexité du modèle q. En effet, la log-vraisemblance maximale doit se comporter de manière linéaire par rapport à la complexité des grands modèles dont le biais s'annule, les modèles plus complexes n'apportant que de la variance supplémentaire. Un choix efficace pour la constante κ s'avère être deux fois la pente fournie par ce comportement linéaire. Une méthodologie équivalente consiste à considérer le double de la pénalité entraînant les changements les plus abrupts dans la complexité du modèle sélectionné (voir Arlot (2019) pour une revue complète de ces méthodes et Baudry (2009) pour une application aux mélanges).

Cette méthode peut s'avérer très utile dans le cas d'une modélisation mal spécifiée ou quand l'asymptotique n'est pas justifiée. C'est en particulier le cas dans l'application présentée dans la section suivante.

1.4 Une application : carte d'identité d'une tumeur

Afin d'illustrer l'intérêt des modèles de mélange dans les applications, je présente ici un travail récent réalisé en collaboration avec Yves Rozenholc et Tatiana Popova, mettant en œuvre un modèle de mélange gaussien dont les centres sont contraints à appartenir à une structure dont certaines caractéristiques sont à inférer (Liu et al., 2015; Keribin et al., 2019).

La caractérisation des altérations du nombre de copies dans le génome est d'importance capitale pour développer une médecine personnalisée en cancérologie. Le génotype de la tumeur est caractérisé par le nombre total d'allèles, c'est-à-dire le nombre de copies (cn), et la fréquence d'allèles B (baf) pour le déséquilibre allélique. L'échantillon tumoral analysé représente généralement un mélange de cellules tumorales et normales. Soit p la proportion d'ADN normal dans l'échantillon, on a :

$$\mathrm{cn} = 2p\ + (1-p)(n_A^t + n_B^t) = 2p + (1-p)(u+v),$$

$$\mathrm{baf} = \frac{p\, n_B^s + (1-p)\, n_B^t}{2p\ + (1-p)(n_A^t + n_B^t)}.$$

où n_A^t et n_B^t désignent le nombre de copies d'allèle A et B dans la tumeur, et n_A^s et n_B^s ceux provenant de la partie saine du prélèvement. Les puces à SNPs (*Single Nucleotide Polymorphism*), une variante de puce à ADN, sont encore utilisées pour mesurer les profils d'altération du nombre de copies. En chaque site présentant un polymorphisme (appelé site SNP), elles fournissent directement le signal baf ; le signal cn est obtenu sous la forme

$$\mathrm{lrr} = \alpha \log_2 \mathrm{cn} + \beta,$$

où α est un facteur de contraction dépendant de la plate-forme de micro-array et de conditions expérimentales. C'est pourquoi il n'est pas possible d'en avoir un étalonnage général. Le paramètre β est un décalage constant dû à une ploïdie tumorale inconnue (nombre de jeux complets de chromosomes dans une cellule), et impossible à calculer sans référence à un échantillon entièrement sain.

La méthode GAP (Genome Alteration Print) de Popova et al. (2009), basée sur une segmentation préliminaire des profils (baf,lrr) issus de puces SNPs (figure 1.2), utilise une approche déterministe pour déterminer le profil du nombre absolu de copies. Elle s'appuie

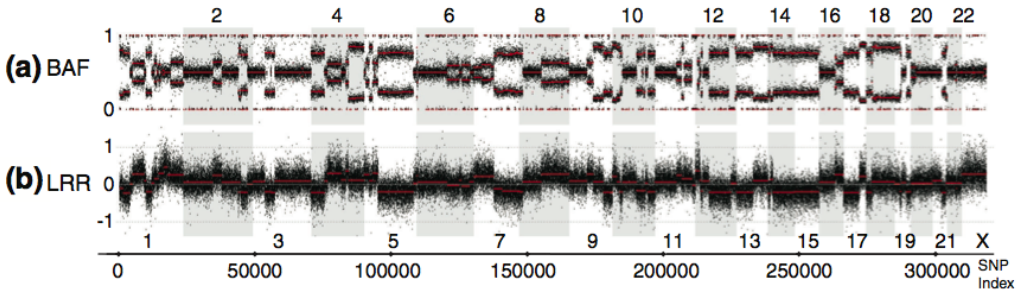


FIGURE 1.2 – Exemple de signaux segmentés baf et lrr, Popova et al. (2009)

sur le fait que les observations représentées dans le plan 2D sont centrées autour de centres contraints d'appartenir à un réseau (figure 1.3) dépendant de paramètres inconnus (p, α, β).

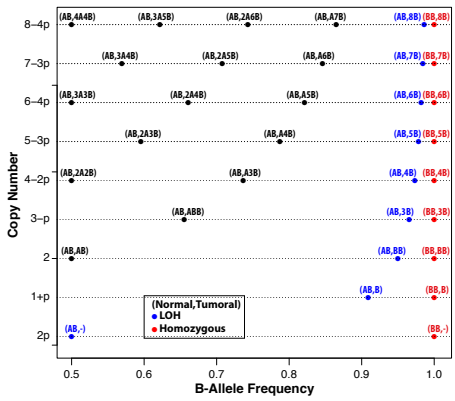


FIGURE 1.3 – Correspondance entre les mutations tumorales et les valeurs théoriques de $(\mathrm{baf}, \mathrm{lrr})$, en supposant que $0 \leq u \leq v$.

Nous avons développé un modèle probabiliste pour la méthode GAP en définissant un modèle de mélange gaussien dont les centres sont contraints d'à appartenir à cette grille :

$$\begin{aligned}
\mathrm{BAF}_i^0 &= \mathrm{baf}_{k(i)}^0 + \frac{\sigma}{\sqrt{n_i^0}}\varepsilon_i^0, \\
\mathrm{BAF}_i^1 &= \mathrm{baf}_{k(i)}^1 + \frac{\sigma}{\sqrt{n_i^1}}\varepsilon_i^1, \quad \text{avec } \mathrm{baf}_{k(i)}^1 = 1, \\
\mathrm{LRR}_i &= \mathrm{lrr}_{k(i)} + \frac{\eta}{\sqrt{n_i}}\xi_i, \quad \text{avec } \mathrm{lrr}_{k(i)} = \alpha \log_2(\mathrm{cn}_{k(i)}) + \beta,
\end{aligned}$$

Les observations sont les moyennes des signaux sur chaque segment : BAF_i^0 fait référence à la moyenne sur un segment de mutation constante du signal BAF des sites hétérozygotes et BAF_i^1 à celle du signal homozygote. Le signal LRR_i est commun sur ce segment, et moyenné sur sa longueur $n_i = n_i^0 + n_i^1$. L'estimation est effectuée à l'aide d'un algorithme EM permettant d'accéder non seulement aux paramètres mais aussi au nombre altéré de copies le plus probable sur chaque segment ainsi que la proportion tumorale inconnue.

La sélection de modèle par critère BIC (section 1.3.1) a bien fonctionné sur les données simulées, retournant le nombre de copies maximal adéquat, même pour une forte proportion p de tissu sain, mais il sous-pénalise nettement quand il est appliqué sur des

données de cancer du côlon. L'heuristique de pente (section 1.3.3) a retrouvé la bonne valeur du nombre de copies maximales. De fait, la constante κ estimée adaptativement sur les données (figure 1.4) est supérieure à la pénalité du critère BIC. En dehors du fait que le

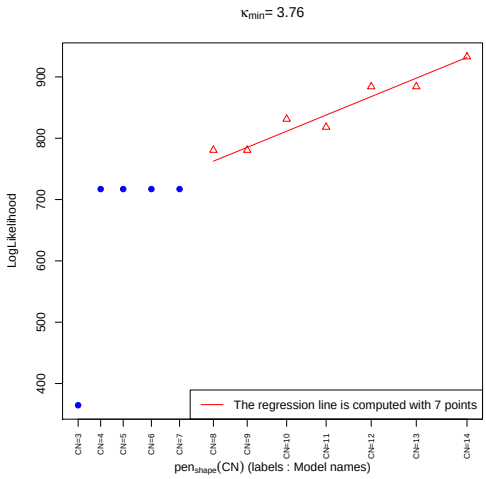


FIGURE 1.4 – Heuristique de pente pour les données de cancer du côlon.

modèle de mélange ne peut être qu'une grossière approximation du phénomène observé et donc induire BIC à surparamétrer, il est à noter ici que les termes d'ordre 1 en probabilité ne sont pas pas négligeables, du fait des données moyennées sur des segments relativement longs : environ 200 segments de longueur entre 500 à 1000 SNPs. Ainsi l'utilisation de l'heuristique de pente lorsque le maximum de vraisemblance présente une tendance linéaire pour les grandes valeurs de complexité est un moyen efficace de sélectionner un modèle. Dans tous les cas, lorsque la pénalité de BIC est inférieure à la pénalité minimale, BIC ne devrait pas être utilisé, ou du moins son approximation devrait être reconsidérée (Keribin, 2019). Enfin, cet exemple montre un cas où l'heuristique de pente fonctionne bien avec un biais de modèle. On pourrait d'ailleurs supposer que l'heuristique de pente peut être justifiée lorsqu'un biais existe et reste constant pour les grands modèles. L'estimation et la sélection permettent de définir une carte d'identité caractérisant la tumeur.

1.5 Le clustering des variables pour réduire la dimension

Nous avons jusqu'à présent considéré le clustering des individus. Mais le clustering peut également s'appliquer aux variables. C'est le cas par exemple en imagerie par résonance magnétique fonctionnelle où le nombre de variables (voxels) confronte le décodage du cerveau à la grande dimension. Ce décodage étant une tâche supervisée, et la neuro-imagerie une application importante, leurs présentations nécessitent un chapitre 4 spécifique, dans lequel sera illustré l'apport du clustering de variables à un apprentissage supervisé en grande dimension.

Chapitre 2

Co-clustering et modèle des blocs latents

Cette section rapporte les travaux faits sur le modèle des blocs latents. Initiés par une collaboration avec Gilles Celeux et Gérard Govaert (Keribin et al., 2010), l'étude des algorithmes d'estimation (Keribin et al., 2015) a été développée et approfondie pendant la thèse de Vincent Brault, co-dirigée avec Gilles Celeux. L'étude de la consistance et normalité asymptotique des estimateurs du maximum de vraisemblance et variationnel a été entreprise après la thèse en collaboration avec Vincent Brault et Mahendra Mariadassou (Brault et al., 2019).

2.1 Coclustering

Le clustering permet de partitionner une matrice d'observations $\mathbf{x} = (x_{ij})$ à n lignes et d colonnes. Les lignes peuvent par exemple être vues comme des individus et les colonnes comme des variables.

Dans de nombreuses applications, et en particulier quand le nombre de variables est important, il est également intéressant d'effectuer un clustering des colonnes de façon simultanée à celui des individus : l'objectif est la détection de blocs d'observations homogènes mais dissemblables entre eux. C'est le principe du co-clustering qui est illustré sur la figure 2.1 avec une matrice d'observations binaires : le clustering permet de passer du tableau initial (1) au tableau dont les lignes ont été réordonnées suivant leur clustering (2), puis les colonnes réordonnées suivant les clusters-colonne faisant apparaître la structure de blocs (3). L'information du tableau initial peut être résumée par la matrice (4) : la représentation ainsi obtenue est très parcimonieuse.

Le co-clustering peut avoir de nombreuses applications, par exemple :

— systèmes de recommandation : déterminer des groupes de consommateurs achetant préférentiellement des groupes de biens de consommation (une cellule contient une valeur binaire) et réciproquement ; groupes de consommateurs notant de façon similaire des groupes de biens de consommation (une cellule contient une valeur entière, données ordinales).

— analyse d'expression de gènes : grouper des gènes qui s'expriment de façon similaire suivant des groupes de conditions de stress (et réciproquement), afin de détecter des gènes orphelins. Une cellule contient une donnée numérique continue.

— analyse de texte : déterminer, à partir de la table de contingence documents/mots, des groupes de documents qui contiennent des profils similaires de groupes de mots

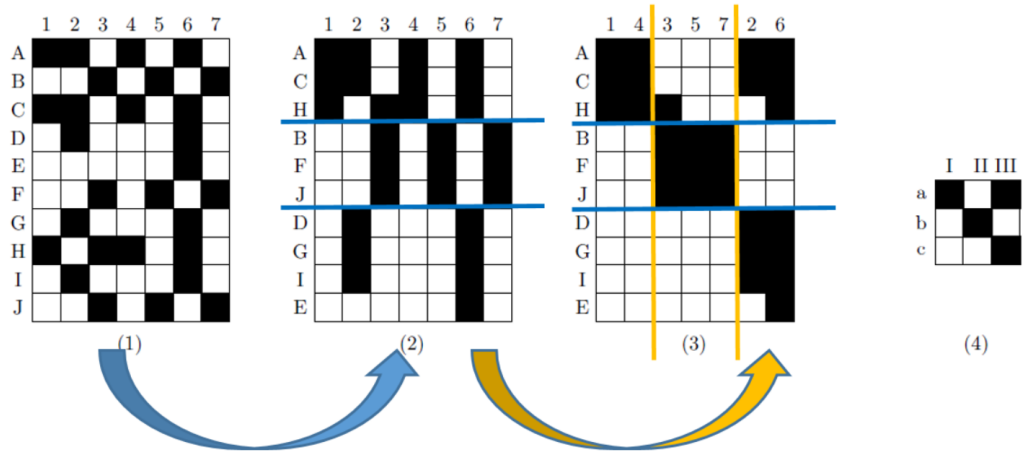


FIGURE 2.1 – Du clustering au co-clustering (à partir de Govaert et Nadif (2008))

et réciproquement. Une cellule contient une donnée de comptage.

Le co-clustering est l'identification non supervisée de la structure croisée précédente, dont la forme est définie par l'hypothèse suivante :

H_1 : produit cartésien *Les blocs sont le résultat du produit cartésien d'une partition (latente) des lignes en g clusters-ligne par une partition (latente) des colonnes en m clusters-colonne.*

On introduit ainsi $\mathbf{z} = (z_{ik})_{i=1,\dots,n,k=1,\dots,g}$, matrice de classification des lignes où $z_{ik} = 1$ si la ligne i appartient au bloc-ligne k et $z_{ik} = 0$ sinon ; $\mathbf{w} = (w_{j\ell})_{j=1,\dots,d,k=1,\dots,g}$, matrice de classification des colonnes où $w_{j\ell} = 1$ si la colonne j appartient au bloc-colonne ℓ et $w_{j\ell} = 0$ sinon.

L'hypothèse H_1 exclut les cas de chevauchement ou de non alignement de blocs, et nécessite un partitionnement complet de la matrice, contrairement au *bi-clustering*, qui peut ne rechercher qu'un bloc de caractéristiques homogènes.

Si sur l'exemple de la figure 2.1, il semble simple d'arriver à une solution, ceci n'est pas le cas de façon générale dès que la taille de la matrice et le nombre de groupes augmentent. Il existe de nombreuses méthodes de co-clustering qu'elles soient déterministes (basées sur la *reconstruction*) ou basées sur une modélisation probabiliste (*model-based*), voir Brault et Lomet (2015) pour une revue bibliographique récente :

- les méthodes déterministes comprennent l'utilisation de mesures de dissimilarité (Govaert, 1977, 1995; Banerjee et al., 2007), la factorisation de matrices non négatives (Lee et Seung, 2001), la décomposition en blocs de valeurs non-négatives (Long et al., 2005) ou la tri-factorisation orthogonale de matrice non-négative (Yoo et Choi, 2010)
- les modèles probabilistes (*model-based*), dont les modèles de mélange et leurs variantes, utilisent des variables latentes pour définir les classes en ligne et en colonne (Govaert et Nadif, 2003; Shan et Banerjee, 2008; Wyse et Friel, 2012).

2.2 Hypothèses du modèle des blocs latents

Le modèle des blocs latents (*latent block model*, LBM) définit un modèle de co-clustering vérifiant H_1 avec le modèle génératif d'hypothèses suivantes (Govaert et Nadif, 2008, 2013) :

H_2 : indépendance des labels *Les variables latentes* $\mathbf{z}$ *et* $\mathbf{w}$ *sont indépendantes. Les labels-ligne* $\mathbf{z}$ *sont i.i.d. de loi multinomiale*

$$z_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$$

avec $\pi_k = \mathbb{P}(z_{ik} = 1)$ *pour* $k = 1, \ldots, g$ *et* $i = 1, \ldots, n$. *Les labels-colonne* $\mathbf{w}$ *sont i.i.d. de loi multinomiale*

$$w_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_m))$$

avec $\rho_\ell = \mathbb{P}(w_{j\ell} = 1)$ *pour* $\ell = 1, \ldots, m$ *et* $j = 1, \ldots, d$

H_3 : indépendance conditionnelle *Les observations* (x_{ij}), *conditionnellement aux labels* $(\mathbf{z}, \mathbf{w})$, *sont des variables aléatoires indépendantes dont la loi appartient à une famille paramétrique* $\mathcal{F}$; *la densité conditionnelle de* x_{ij} *dépend uniquement uniquement de son bloc d'appartenance* (k, ℓ) :

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

Ce modèle peut être décliné dans de nombreuses versions suivant le type de données manipulées : binaires (Govaert et Nadif, 2008), gaussiennes (Lomet, 2012), multinomiales (Keribin et al., 2015) ou poissoniennes (Govaert et Nadif, 2010; Robert, 2017) par exemple.

Notation La figure 2.2 issue de Brault (2014) présente schématiquement les notations utilisées.

LBM et SBM Le modèle stochastique de blocs (SBM, Daudin et al. (2008)) est un modèle voisin du LBM, pour lequel $\mathbf{z} = \mathbf{w}$, et la matrice $\mathbf{x}$ représente la matrice d'adjacence du graphe des relations entre des individus. Dans ce modèle, les sommets sont échantillonnés dans une population et la préoccupation concerne les paramètres de population, c'est-à-dire les poids de chaque classe et leurs paramètres de connectivité. La double structure cachée du LBM rend plus délicate l'étude du modèle, en particulier celle de l'asymptotique qui doit considérer deux directions distinctes en ligne et colonne.

2.3 Défis de la vraisemblance

Sous ces hypothèses, la vraisemblance s'écrit

$$\begin{aligned} p(\mathbf{x}; \theta) &= \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}, \mathbf{w}) p(\mathbf{x}|\mathbf{z}, \mathbf{w}) \\ &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}} \end{aligned} \tag{2.1}$$

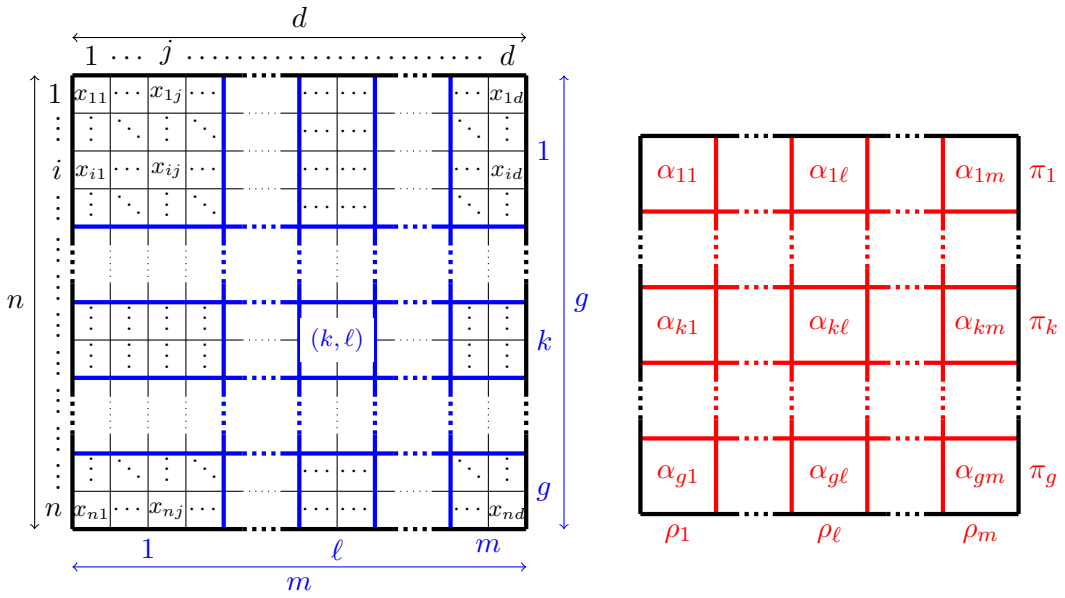


FIGURE 2.2 – Récapitulatif des notations utilisées

où $\mathcal{Z} \times \mathcal{W}$ représente l'ensemble de toutes les partitions croisées de $n \times d$ cellules en $g \times m$ blocs. Le paramètre du modèle est

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \Pi_g \times \Pi_m \times \Xi^{gm} := \boldsymbol{\Theta}$$

en reprenant les notations définies à la section 1.2. Il permet de réduire drastiquement la dimension par rapport à un modèle de mélange simple, comme l'illustre les courbes de la figure 2.3.

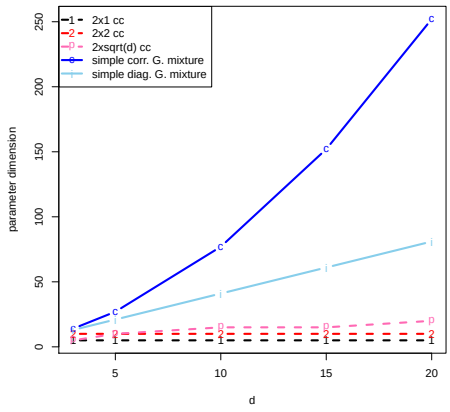


FIGURE 2.3 – Évolution, en fonction du nombre de variables, de la dimension d'un modèle de mélange gaussien de variance diagonale, d'un modèle de mélange gaussien à variables corrélées, d'un modèle des blocs latents avec un nombre de clusters en colonne de 1, 2 ou la racine carrée du nombre de variables. Le nombre de clusters en ligne est identique dans tous les cas.

Le LBM peut être vu comme un *modèle de mélange généralisé* : en effet, l'expression (2.1) ne peut se factoriser à cause de la dépendance complexe qu'induit la double structure

$(\mathbf{z}, \mathbf{w})$ sur les observations (x_{ij}), contrairement à son expression analogue pour les modèles de mélange simple

$$p(\mathbf{x};\theta) = \sum_{z\in\mathcal{Z}} \prod_i \pi_k^{z_{ik}} \prod_{i,k} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}} = \prod_i \left(\sum_k \pi_k \varphi(x_{ij};\alpha_{k\ell}) \right).$$

L'expression ne se factorisant pas, le calcul de la vraisemblance du LBM ou de son logarithme nécessite la somme de $g^n m^d$ termes, ce qui n'est pas réalisable en temps raisonnable avec les moyens informatiques existants, même pour un faible nombre d'observations et de blocs. Ce point sera traité dans la section 2.5. Mais avant de considérer le problème de l'estimation, il est important d'analyser l'identifiabilité du modèle.

2.4 Identifiabilité

Rappelons que le mélange simple de lois de Bernoulli multivariée n'est pas identifiable (Gyllenberg et al., 1994), indépendamment du problème de permutation des labels, mais Allman et al. (2009) donne un ensemble de conditions suffisantes pour leur identifiabilité. Le théorème suivant définit un ensemble de conditions suffisantes garantissant l'identifiabilité du LBM binaire :

Théorème 2 (Identifiabilité du LBM binaire, Keribin et al. (2015))**.** *Soient $\boldsymbol{\pi}$ (resp. $\boldsymbol{\rho}$) les poids de mélange des lignes (resp. colonnes) d'un LBM binaire et soit $A = (\alpha_{kl})$ ma matrice de dimension $g \times m$ des paramètres des lois de Bernoulli. On suppose que :*

- *C_1 : pour tout $1 \leq k \leq g$, $\pi_k > 0$ et les coordonnées du vecteur $\boldsymbol{\tau} = A\boldsymbol{\rho}$ sont distinctes.*
- *C_2 : for all $1 \leq \ell \leq m$, $\rho_\ell > 0$ et les coordonnées du vecteur $\boldsymbol{\sigma} = \boldsymbol{\pi}'A$ sont distinctes (où $\boldsymbol{\pi}'$ est la transposée $\boldsymbol{\pi}$).*

alors, le LBM est identifiable pour $n \geq 2m-1$ et $d \geq 2g-1$.

Les conditions C_1 et C_2 sont des extensions de celles permettant de garantir l'identifiabilité du modèle SBM (Celisse et al., 2012), où $n = d$ et $\mathbf{z} = \mathbf{w}$. Elles sont peu restrictives puisqu'elles sont vérifiées sauf sur un ensemble de mesure nulle : on dit que le modèle est *génériquement identifiable*. Elles imposent que les probabilités $\tau_k = \mathbb{P}(y_{ij} = 1 | z_{ik} = 1)$ (resp. $\sigma_\ell = \mathbb{P}(y_{ij} = 1 | w_{j\ell} = 1)$) puissent être triées en ordre strictement croissant, permettant de les utiliser pour définir un ordre naturel sur les classes.

Le principe de la preuve repose sur l'utilisation de la matrice de Vandermonde des coefficients $(\tau_k)^i$ pour $0 \leq i < g$ et $1 \leq k \leq g$ et l'étude de la matrice des probabilités d'obtenir une valeur 1 sur les i premières cellules de la première ligne de $\mathbf{x}$. Ces résultats s'étendent aisément au LBM à observations multinomiales, et à toute loi dont les paramètres peuvent être identifiés à partir d'une discrétisation de la loi, ce qui est vérifié pour toutes les grandes lois classiques, comme les familles exponentielles univariées.

2.5 Algorithmes

La vraisemblance ne peut être calculée numériquement en temps raisonnable même pour un faible nombre d'observations et un faible nombre de blocs : par exemple, elle demande le calcul de 10^{12} termes non factorisables pour une matrice de taille 10×10 sur laquelle sont définis 2×2 blocs.

De plus, le calcul de l'espérance de la vraisemblance complète conditionnellement aux observations

$$Q(\theta|\theta^{(c)}) = \sum_{i,k} t_{ik}^{(c)} \log \pi_k + \sum_{j,\ell} s_{j\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} e_{i,j,k,\ell}^{(c)} \log \varphi(x_{ij}; \alpha_{k\ell}) \tag{2.2}$$

nécessite le calcul des lois des labels conditionnellement aux observations

$$t_{ik}^{(c)} = \mathbb{P}(z_{ik} = 1|\mathbf{x}; \theta^{(c)}), \quad s_{j\ell}^{(c)} = \mathbb{P}(w_{j\ell} = 1|\mathbf{x}; \theta^{(c)})$$

et

$$e_{i,j,k,\ell}^{(c)} = \mathbb{P}(z_{ik} w_{j\ell} = 1|\mathbf{x}; \theta^{(c)}).$$

Ces termes nécessitent à leur tour le calcul d'un nombre exponentiel d'opérations non factorisables. Ainsi, l'étape E de l'algorithme EM doit être aménagée.

2.5.1 VEM et SEM-Gibbs

Dans Keribin et al. (2010), nous avons comparé deux méthodes permettant de contourner le problème : le calcul d'un estimateur variationnel (VEM) comme approximation de l'estimateur du maximum de vraisemblance et l'adaptation d'un EM stochastique (SEM-Gibbs).

L'approximation variationnelle de l'estimateur du maximum de vraisemblance (cf section1.2.3) découle de l'approximation faite à l'étape E, où l'énergie libre

$$\mathcal{F}(q_{wz}) = \mathbb{E}_{q_{wz}} \left(\log \frac{p(\mathbf{x}, \mathbf{w}, \mathbf{z}; \theta^{(c)})}{q_{wz}(\mathbf{w}, \mathbf{z})} |\mathbf{x} \right)$$

est maximisée en se restreignant aux lois libres factorisées :

$$q_{zw}(\mathbf{z}, \mathbf{w}) \simeq q_z(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(c)}) q_w(\mathbf{w}|\mathbf{x}; \boldsymbol{\theta}^{(c)}) = \prod_i q_{z_i}(z_i|\mathbf{x}; \boldsymbol{\theta}^{(c)}) \prod_j q_{w_j}(w_j|\mathbf{x}; \boldsymbol{\theta}^{(c)})$$

Les calculs deviennent aisés avec cette approximation. Ceci est équivalent à supposer que $e_{i,j,k,\ell} = t_{ik} s_{j\ell}$, méthode utilisée par Govaert et Nadif (2008), menant à l'algorithme BEM, que nous avons appelé VEM pour mettre l'accent sur l'interprétation variationnelle. L'estimateur variationnel est défini par

$$\widehat{\boldsymbol{\theta}}_{VAR} = \arg \max_{\boldsymbol{\theta}, q_z, q_w} \mathcal{F}(q_z, q_w, \boldsymbol{\theta})$$

Algorithme 4 (VEM, Govaert et Nadif (2008))**.** *Après initialisation, itérer :*

1. *Étape E : Maximisation de l'énergie libre jusqu'à convergence*
   (a) *calcul de s_{ik} à t_{jl} et $\theta^{(c)}$ donnés*
   (b) *calcul de t_{jl} à s_{ik} et $\theta^{(c)}$ donnés*
   $\hookrightarrow$ *d'où $s^{(c+1)}$ et $t^{(c+1)}$*
2. *Étape M : mise à jour de $\theta^{(c+1)}$*

Une variante permet de ne faire qu'une itération dans l'étape E. Si l'algorithme VEM permet le calcul de l'estimateur variationnel, il possède deux inconvénients :

— VEM est fortement dépendant des conditions initiales.

— L'estimateur variationnel résultant, tout au moins à distance finie, n'est qu'une approximation de l'estimateur du maximum de vraisemblance. En effet, d'après Gunawardana et Byrne (2005), 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.

L'étude des propriétés asymptotiques de ces estimateurs s'est révélée ardue à cause de la dépendance complexe des observations apportée par le modèle de blocs et n'a été que récemment résolue, cf section 2.6.

Afin de contourner le problème d'initialisation, nous avons proposé d'utiliser une version stochastique de l'EM (SEM, Celeux et Diebolt (1986); Celeux et al. (1996)), dans laquelle l'étape d'estimation est remplacée par la génération d'un échantillon des données manquantes $(\mathbf{w}^{(c)}, \mathbf{z}^{(c)})$ sous la loi des données manquantes conditionnellement aux observations et à l'état en cours $\theta^{(c)}$ du paramètre : on obtient ainsi un pseudo-échantillon complet. L'étape de maximisation recherche le paramètre maximisant la vraisemblance complétée, dans laquelle les variables manquantes sont remplacées par leur tirage.

La loi $p(\mathbf{w}, \mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(c)})$ est toujours impossible à calculer, mais les lois $p(\mathbf{w}|\mathbf{x}, \mathbf{z}; \theta^{(c)})$ et $p(\mathbf{z}|\mathbf{x}, \mathbf{w}; \theta^{(c)})$ sont facilement calculables. Par exemple,

$$p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(t)}; \theta^{(c)}) = \prod_i p(z_i|x_{i\cdot}, \mathbf{w}^{(c)})$$

et

$$p(z_i = k|x_{i\cdot}, \mathbf{w}^{(c)}) = \frac{\pi_k \psi_k(x_{i\cdot}, \alpha_{k\cdot})}{\sum_{k'} \pi_{k'} \psi_{k'}(x_{i\cdot}, \alpha_{k'\cdot})}, k = 1, \dots g$$

où $x_{i\cdot}$ désigne la i^{e} ligne de la matrice $\mathbf{x}$, $\alpha_{k\cdot} = (\alpha_{k1}, \dots, \alpha_{km})$ et

$$\psi_k(x_{i\cdot}, \alpha_{k\cdot}) = \prod_\ell \alpha_{k\ell}^{u_{i\ell}} (1 - \alpha_{k\ell})^{d_\ell - u_{i\ell}}, \quad u_{i\ell} = \sum_j w_{j\ell}^{(c)} x_{ij}, \quad d_\ell = \sum_j w_{j\ell}^{(c)}.$$

Ceci permet l'utilisation d'un échantillonneur de Gibbs, itérant le tirage de $\mathbf{w}$ suivant $p(\mathbf{w}|\mathbf{x}, \mathbf{z}; \theta^{(c)})$, puis celui de $\mathbf{z}$ suivant $p(\mathbf{z}|\mathbf{x}, \mathbf{w}; \theta^{(c)})$. SEM-Gibbs est donc l'itération successive de deux étapes, la première étant elle-même une itération d'un schéma de Gibbs pour simuler les données manquantes :

Algorithme 5 (SEM-Gibbs)**.** *Après initialisation, itérer :*

1. *Étape E-S : Répéter les deux étapes suivantes*
   (a) *estimation puis tirage de* $\mathbf{z}^{(t+1)}$ *suivant la loi* $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(t)}; \theta^{(c)})$
   (b) *estimation puis tirage de* $\mathbf{w}^{(t+1)}$ *suivant la loi* $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(t+1)}; \theta^{(c)})$ *(expression analogue en remplaçant les lignes par les colonnes)*
   (c) *d'où* $w^{(c+1)}$ *et* $z^{(c+1)}$
2. *Étape M : mise à jour de* $\theta^{(c+1)}$

VEM est basé sur une approximation numérique. SEM-Gibbs n'utilise aucune approximation, les lois utilisées dans les différentes étapes de l'échantillonneur de Gibbs étant exactes. En revanche, SEM (et donc SEM-Gibbs) n'augmente pas la probabilité vraisemblance à chaque itération, mais génère une chaîne de Markov irréductible avec une distribution stationnaire unique qui doit être concentrée autour de l'estimation du paramètre ML (McLachlan et Krishnan, 2007). Ainsi, une estimation naturelle de $\boldsymbol{\theta}$ dérivée de SEM-Gibbs est la moyenne des valeurs des $\boldsymbol{\theta}^{(c)}$ obtenues après une période de rodage.

Les simulations (Keribin et al., 2012) ont montré que l'algorithme SEM-Gibbs permet de s'affranchir raisonnablement des problèmes d'initialisation préconisant d'initialiser par SEM-Gibbs un algorithme VEM.

2.5.2 Régulariser par l'inférence bayésienne : EM-VBayes et Gibbs

La méthodologie précédente peut donner des estimations satisfaisantes mais elle a une tendance marquée à fournir des clusters vides, et à produire moins de clusters que le vrai modèle de simulation après une règle d'affectation du MAP. Le phénomène est illustré sur la figure 2.4.

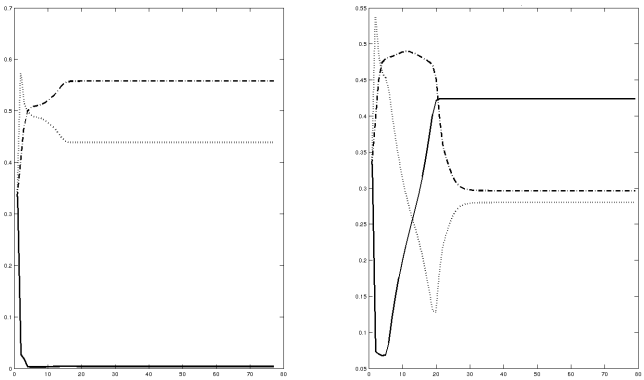


FIGURE 2.4 – Illustration du phénomène de dégénéresence pour un LBM avec trois classes en ligne : l'estimation des poids des clusters suivant les algorithmes VEM (à gauche) et VBayes (à droite), les deux algorithmes partant de la même initialisation

Ce point méthodologique a été traité dans la thèse de VincentBrault (2014) où le cadre des données binaires est étendu à celui des données qualitatives où la loi de l'observation x_{ij} du bloc (k,ℓ) est une multinomiale $\mathcal{M}(1;\alpha_{k\ell})$, de paramètre $\alpha_{k\ell} = (\alpha^h_{k\ell})_{h=1,\dots,r}$, où $\alpha^h_{k\ell} \in (0;1)$ et $\sum_h \alpha^h_{k\ell} = 1$ et de densité

$$\varphi(\mathbf{x}_{ij};\alpha_{k\ell}) = \prod_h (\alpha^h_{k\ell})^{x^h_{ij}}$$

L'inférence bayésienne, utilisée comme un outil de *régularisation*, s'y est révélée efficace pour atténuer le problème de dégénérescence des classes. Pour le LBM multinomial, il existe des distributions a priori appropriées non informatives, pour les proportions de mélange $\boldsymbol{\pi}$ et $\boldsymbol{\rho}$, et pour le paramètre $\boldsymbol{\alpha}$:

$$\boldsymbol{\pi} \sim \mathcal{D}(a,\dots,a), \quad \boldsymbol{\rho} \sim \mathcal{D}(a,\dots,a), \; \alpha_{k\ell} \sim \mathcal{D}(b,\dots,b), \tag{2.3}$$

où $\mathcal{D}(v,\dots,v)$ représente une loi de Dirichlet de paramètre v. Le modèle graphique est présenté sur la figure 2.5.

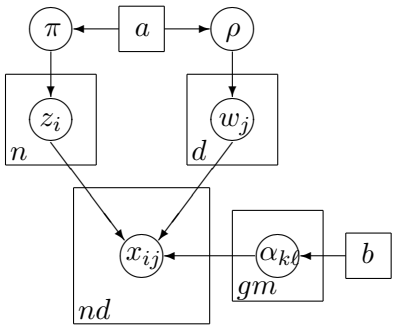


FIGURE 2.5 – Représentation graphique du LBM bayésien

Dans (Brault et al., 2012; Keribin et al., 2015), nous avons adapté l'algorithme 3 EM-VBayes au LBM. L'étape E est la même que celle de VEM (algorithme 4), seule la maximisation diffère

$$\pi_k^{(c+1)} = \frac{a - 1 + s_{.k}^{(c+1)}}{n + g(a-1)}, \quad \rho_\ell^{(c+1)} = \frac{a - 1 + t_{.\ell}^{(c+1)}}{d + m(a-1)}$$

$$\alpha_{k\ell}^{h\ (c+1)} = \frac{b - 1 + \sum_{i,j} s_{ik}^{(c+1)} t_{j\ell}^{(c+1)} x_{ij}^h}{r(b-1) + s_{.k}^{(c+1)} t_{.\ell}^{(c+1)}},$$

où $s_{ik}^{(c+1)}$ et $t_{j\ell}^{(c+1)}$ sont les valeurs courantes des lois conditionnelles des labels, et $s_{.k}^{(c+1)}$ et $t_{.\ell}^{(c+1)}$, leurs sommes sur les lignes et colonnes respectivement. La régularisation vient des hyperparamètres a et b et leur choix est important : on retrouve VEM quand $a = b = 1$, et EM-VBayes a tendance à produire plus de classes vides avec l'a priori de Jeffreys ($a = b = 1/2$). Les simulations ont montré que prendre $a > 1$ avait un impact bénéfique pour éviter la dégénérescence des classes. Ceci va dans le même sens que les résultats de Frühwirth-Schnatter (2006) pour les modèles de mélange simple ($a = 4$ pour des dimensions modérées et $a = 16$ pour de plus grandes dimensions). Au contraire, prendre $b > 1$ est préjudiciable. Ceci pourrait s'expliquer par le fait que l'a priori sur les lois conditionnelles met plus de poids sur des paramètres de proportion égaux, et donc pénaliserait les solutions qui apporteraient une séparation des blocs.

Si EM-VBayes permet d'éviter la dégénérescence des classes, il est toutefois sensible à l'initialisation comme tout algorithme bayésien variationnel. Ainsi, de même que SEM-Gibbs a permis de définir une initialisation pour VEM, utiliser un échantillonnage de Gibbs est une bonne stratégie pour initialiser un algorithme EM-VBayes, comme l'ont illustré des expérimentations. Celui-ci est facilement mis en place grâce aux propriétés de conjugaison des lois a priori utilisées Brault et al. (2014).

Algorithme 6 (Full Gibbs)**.** *Après une initialisation, itérer*

1. *simuler* $\mathbf{z}^{(c+1)} \sim p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)})$ *comme pour SEM-Gibbs*
2. *simuler* $\mathbf{w}^{(c+1)} \sim p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c+1)}; \boldsymbol{\theta}^{(c)})$ *comme pour SEM-Gibbs*
3. *simuler* $\boldsymbol{\pi}^{(c+1)} \sim \mathcal{D}(a + z_{.1}^{(c+1)}, \ldots, a + z_{.g}^{(c+1)})$
   où $z_{.k} = \sum_i z_{ik}$ *est le nombre de lignes du cluster-ligne* k
4. *simuler* $\boldsymbol{\rho}^{(c+1)} \sim \mathcal{D}(a + w_{.1}^{(c+1)}, \ldots, a + w_{.m}^{(c+1)})$
   où $w_{.\ell} = \sum_j w_{j\ell}$ *le nombre de colonnes du cluster-colonne* ℓ
5. *simuler* $\boldsymbol{\alpha}_{k\ell}^{(c+1)} \sim \mathcal{D}(b + N_{k\ell}^{1\ (c+1)}, \ldots, b + N_{k\ell}^{r\ (c+1)})$
   pour $k = 1, \ldots, g; \ell = 1, \ldots, m$ *et avec*

$$N_{k\ell}^{h\ (c+1)} = \sum_{i,j} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} x_{ij}^h. \tag{2.4}$$

2.6 Consistance et normalité asymptotique

L'étude théorique des propriétés asymptotiques (consistance, loi asymptotique) de l'estimteur du maximum de vraisemblance ou de l'estimateur variationel est un problème délicat. Des résultats partiels existaient pour le LBM et pour le SBM. Celisse et al. (2012)

ont montré pour le SBM (théorème 3) que, sous la vraie valeur du paramètre, la loi des labels conditionnellement aux observations tend vers un Dirac de support les vrais labels. Cette convergence est de plus valide sous la valeur estimée du paramètre si l'estimateur du paramètre de la loi conditionnelle converge à une vitesse d'au moins n^{-1} vers la vraie valeur, où n est le nombre de nœuds du graphe (proposition 3.8). Cette hypothèse n'est pas anodine, et il n'était pas établi qu'un tel estimateur existe sauf dans certains cas particuliers (Ambroise et Matias, 2012). Mariadassou et Matias (2015) ont présenté un cadre unifiant SBM et LBM pour des observations de familles exponentielles, et montré la convergence de la loi conditionnelle des labels pour toute valeur du paramètre dans un voisinage de la vraie valeur pour des observations statisfaisant une à une inégalité de concentration. Bickel et al. (2013) ont prouvé la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle SBM binaire. Rompant avec la vision des précédents auteurs, ils ont d'abord étudié le comportement asymptotique de l'estimateur du maximum de vraisemblance dans le modèle complet (observations et labels) qui est plus simple à manipuler, puis ont prouvé que la vraisemblance complète et la vraisemblance des observations avaient des comportements asymptotiques similaires, en utilisant en particulier une inégalité de Bernstein pour des observations bornées. Mais ces auteurs n'ont pas résolu la prise en compte des complications liées aux symétries que peuvent présenter certains modèles.

Nous avons étendu les résultats de Bickel et al. (2013) au LBM pour des observations de familles exponentielles (Brault et al., 2019), tout en traitant le cas des symétries. Soit $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta})$ la log-vraisemblance des observations et $\mathcal{L}_c(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \log p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})$ la log-vraisemblance des données complètes (données observées et labels). La preuve reprend le schéma de celle de Bickel et al. (2013) :

1. Montrer que l'estimateur du maximum de vraisemblance des données complètes $\widehat{\boldsymbol{\theta}}$ est consistant et que le modèle est asymptotiquement localement normal
2. Prouver que $\mathcal{L}_c(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})$ est "proche" de $\mathcal{L}(\boldsymbol{\theta})$
3. Montrer l'équivalence asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_{MLE}$ et de l'estimateur du maximum de la vraisemblance complète $\widehat{\boldsymbol{\theta}}$, leur différence étant en $o_P(1)$.

Ni l'étape 1 (application classique de la théorie asymptotique d'observations i.i.d.), ni l'étape 3 (simple raisonnement par l'absurde) ne posent de difficulté, c'est l'étape 2 qui est délicate et qui donne lieu à notre résultat principal.

La mise en correspondance des estimateurs $\widehat{\boldsymbol{\theta}}$ et $\widehat{\boldsymbol{\theta}}_{MLE}$ amène à travailler avec la vraisemblance complète et nécessite de définir les situations de symétrie. Nous commencerons par les décrire dans la section suivante.

2.6.1 Symétrie et distance

Les modèles de mélange, et donc le LBM, ne sont définis qu'à une permutation près. Ainsi, soit s une permutation sur $\{1, \ldots, g\}$ et t une permutation sur $\{1, \ldots, m\}$. Deux paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\theta}'$ sont *equivalents*, noté $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$, s'ils sont égaux à une permutation près, i.e. s'il existe deux permutations s et t telles que

$$\boldsymbol{\theta}^{s,t} = (\boldsymbol{\pi}^s, \boldsymbol{\rho}^t, \boldsymbol{\alpha}^{s,t}) = (\boldsymbol{\pi}', \boldsymbol{\rho}', \boldsymbol{\alpha}')$$

Alors, dans le modèle observé, $\mathcal{L}(\boldsymbol{\theta}^{s,t}) = \mathcal{L}(\boldsymbol{\theta}')$. Dans le modèle complet, la permutation doit intervenir également sur les labels : si $(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$ et $(\boldsymbol{\theta}', \mathbf{z}', \mathbf{w}')$ sont équivalents si

$$(\boldsymbol{\theta}^{s,t}, \mathbf{z}^s, \mathbf{w}^t) = (\boldsymbol{\theta}', \mathbf{z}', \mathbf{w}')$$

Alors $\mathcal{L}_c(\theta^{s,t}, \mathbf{z}^s, \mathbf{w}^t) = \mathcal{L}_c(\theta', \mathbf{z}', \mathbf{w}')$. En général, $\mathcal{L}_c(\theta^{s,t}, \mathbf{z}, \mathbf{w}) \neq \mathcal{L}_c(\theta, \mathbf{z}, \mathbf{w})$ sauf si $\boldsymbol{\theta}$ présente une *symétrie*, c'est-à-dire s'il est égal à sa permutation (non triviale) : $\boldsymbol{\theta}^{s,t} = (\boldsymbol{\pi}^s, \boldsymbol{\rho}^t, \boldsymbol{\alpha}^{s,t}) = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$

En d'autres termes, si $\boldsymbol{\theta}$ présente une symétrie, l'affectation de probabilité maximale est *non unique* dans le modèle observé et il en existe au moins $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ d'entre elles. Ceci a des implications importantes pour le comportement asymptotique de la vraisemblance dans le modèle observé.

On définit la distance entre deux configurations $\mathbf{z}$ and $\mathbf{z}^\star$ à une équivalence près (et de façon similaire pour $\mathbf{w}$ and $\mathbf{w}^\star$) par

$$\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} = \inf_{\mathbf{z}' \sim \mathbf{z}} \|\mathbf{z}' - \mathbf{z}^\star\|_0$$

où, pour toute matrice $\mathbf{z}$, $\|\cdot\|_0$ est la distance de Hamming définie par

$$\|\mathbf{z}\|_0 = \sum_{i,k} \mathbb{1}\{z_{ik} \neq 0\}.$$

On note $S(\mathbf{z}^\star, \mathbf{w}^\star, r)$ l'ensemble des configurations qui ont un représentant (pour $\sim$) à distance au plus rn de $\mathbf{z}^\star$ et au plus de rd de $\mathbf{w}^\star$.

$$S(\mathbf{z}^\star, \mathbf{w}^\star, r) = \{(\mathbf{z}, \mathbf{w}) : \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \leq rn \text{ et } \|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim} \leq rd\}$$

2.6.2 Résultat principal

Ce résultat est établi pour un LBM dont la loi conditionnelle appartient à une famille exponentielle régulière de dimension 1 mise sous forme canonique et de paramètre $\alpha \in \Xi$:

$$\varphi(x, \alpha) = b(x) \exp(\alpha x - \psi(\alpha)).$$

tel que $\varphi(x, \alpha)$ est bien défini pour tout $\alpha \in \Xi$. On suppose que le modèle LBM vérifie :

A_1 : Il existe une constante positive c et un compact C_α tels que

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

A_2 : Le vrai paramètre $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ appartient à l'intérieur de $\boldsymbol{\Theta}$

A_3 : $\boldsymbol{\theta}^\star$ est identifiable.

L'identifiabilité nécessite l'injectivité de φ en α, et le fait que chaque ligne et chaque colonne de $\boldsymbol{\alpha}^\star$ soit unique. Des conditions suffisantes d'identifiabilité ont été vues à la section 2.4. Puisque nous avons restreint α à un sous ensemble borné de $\mathring{\mathcal{A}}$, il existe deux valeurs positives M_α et κ telles que $C_\alpha + (-\kappa, \kappa) \subset [-M_\alpha, M_\alpha] \subset \mathring{\Xi}$. De plus, il existe $\bar{\sigma} > 0$ et $\underline{\sigma} > 0$ tels que

$$\sup_{\alpha \in [-M_\alpha, M_\alpha]} \mathbb{V}(\mathrm{X}_\alpha) = \bar{\sigma}^2 < +\infty \quad \text{and} \quad \inf_{\alpha \in [-M_\alpha, M_\alpha]} \mathbb{V}(\mathrm{X}_\alpha) = \underline{\sigma}^2 > 0. \tag{2.5}$$

On définit le régime asymptotique suivant, cadrant la vitesse relative en ligne n et colonne d :

A_4 : $n, d \to \infty$ en vérifiant $\log(d)/n \to 0$ et $\log(n)/d \to 0$

Théorème 3 (complete-observed *dans* Brault et al. (2019))**.** *Soit une matrice* $\mathbf{x}$ *de* $n \times d$ *observations provenant d'un LBM d'ordre connu et de loi conditionnelle appartenant à une famille exponentielle régulière de dimension 1.*

Soit $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ *le nombre de couples de permutations* (s,t) *pour lesquels* $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *présente une symétrie et soient* $\mathbf{z}^\star$ *et* $\mathbf{w}^\star$ *les vraies assignations des lignes et colonnes.*

Si les hypothèses A_1 *à* A_3 *sont vérifiées, et sous le régime asymptotique* A_4*, le rapport de vraisemblance observé se comporte comme le rapport de vraisemblance dans le modèle complet, à un facteur multiplicatif près :*

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

où le terme o_P *est uniforme sur tout* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*.*

Remarque Si $\boldsymbol{\theta}$ présente une symétrie, l'attribution de probabilité maximale n'est pas unique dans le modèle observé et $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ termes contribuent de façon identique à la somme. Ceci n'a pas été pris en compte par Bickel et al. (2013), alors que ces contributions devraient également l'être pour le SBM dans le cas de symétries.

Éléments de preuve La preuve travaille conditionnellement à un ensemble d'assignations réelles (et inconnues) $(\mathbf{z}^\star, \mathbf{w}^\star)$ qui ont suffisamment d'observations dans chaque cluster-lignes et cluster-colonnes et appelées régulières : pour tout $c > 0$, $\mathbf{z} \in \mathcal{Z}$ et $\mathbf{w} \in \mathcal{W}$ sont c-*régulières* si

$$\min_k z_{+k} \geq cn \quad \text{et} \quad \min_\ell w_{+\ell} \geq cd. \tag{2.6}$$

Soient $\mathcal{Z}_1$ et $\mathcal{W}_1$ les sous-ensembles de $\mathcal{Z}$ et $\mathcal{W}$ contenant les assignations $c/2$-régulières, avec c défini dans l'hypothèse A_1. Soit Ω_1 l'événement $\{(\mathbf{z}^\star, \mathbf{w}^\star) \in \mathcal{Z}_1 \times \mathcal{W}_1\}$, alors, par l'application d'une simple inégalité de Hoeffding :

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(\bar{\Omega}_1\right) \leq g \exp\left(-\frac{nc^2}{2}\right) + m \exp\left(-\frac{dc^2}{2}\right).$$

L'événement Ω_1 arrive donc avec grande probabilité sous $\mathbb{P}_{\boldsymbol{\theta}^\star}$ dans l'espace $\mathcal{Z} \times \mathcal{W}$, uniformément en $\boldsymbol{\theta}^\star \in \Theta$.

Travaillant conditionnellement à Ω_1, soit $(\mathbf{z}^\star, \mathbf{w}^\star) \in \mathcal{Z} \times \mathcal{W}$ et une séquence t_{nd} tendant vers 0 telle que $t_{nd}^2 \gg \frac{n+d}{nd}$. t_{nd} existe quand $n \to \infty$ et $d \to \infty$, ce que permet l'hypothèse A_4. La preuve repose sur la décomposition suivante de la vraisemblance observée :

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \sim (\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) + \sum_{(\mathbf{z},\mathbf{w}) \nsim (\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \tag{2.7}$$

où le deuxième terme s'avère être négligeable. Notant que la vraisemblance complète peut s'écrire

$$p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}) \exp(F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})).$$

l'étude asymptotique de F_{nd} définie par

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

est cruciale. Son contrôle passe par l'étude du rapport de vraisemblance profilée Λ et son espérance $\tilde{\Lambda}$:

$$\begin{aligned} \Lambda(\mathbf{z},\mathbf{w}) &= \max_{\boldsymbol{\theta}} F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) \\ \tilde{\Lambda}(\mathbf{z},\mathbf{w}) &= \max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}^\star}\left[F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) \middle|\, \mathbf{z}^\star,\mathbf{w}^\star\right] \end{aligned} \tag{2.8}$$

On écrit $F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = F_{nd} - \tilde{\Lambda}(\mathbf{z}, \mathbf{w}) + \tilde{\Lambda}(\mathbf{z}, \mathbf{w})$. En fait, ce ne sont pas deux mais trois comportements qui sont à distinguer pour l'étude de la contribution des différentes assignations :

1. *contrôle global* pour les assignations $(\mathbf{z}, \mathbf{w})$ loin de $(\mathbf{z}^\star, \mathbf{w}^\star)$, c'est à dire telle que $\tilde{\Lambda}(\mathbf{z}, \mathbf{w})$ est de l'ordre de $\Omega(-nd)$, la proposition 2 donne les larges déviations pour $F_{nd} - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})$ prouvant que F_{nd} est aussi de l'ordre de $-\Omega_P(nd)$.
   Ainsi, soit une suite t_{nd} tendant vers 0 telle que $t_{nd}^2 \gg \frac{n+d}{nd}$, avec $n \to \infty$ et $d \to \infty$. Alors, conditionnellement à Ω_1 et pour n, d suffisamment grands tels que $2\sqrt{2nd}t_{nd} \geq gm$, on a :
   $$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{(\mathbf{z}, \mathbf{w}) \notin S(\mathbf{z}^\star, \mathbf{w}^\star, t_{nd})} p(\mathbf{z}, \mathbf{w}, \mathbf{x}; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star)) \tag{2.9}$$

2. *contrôle local* : pour les assignations proches $(\mathbf{z}, \mathbf{w})$ de $(\mathbf{z}^\star, \mathbf{w}^\star)$, $\tilde{\Lambda}(\mathbf{z}, \mathbf{w})$ est de l'ordre de $-\Omega(d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim})$ et un résultat de petites déviations est nécessaire pour contrôler $\Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star)$, cf proposition 3. On peut montrer que, sous l'hypothèse A_4, leurs contributions combinées à la vraisemblance est aussi un o_P de $p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)$:
   $$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{\substack{(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, C) \\ (\mathbf{z}, \mathbf{w}) \not\sim (\mathbf{z}^\star, \mathbf{w}^\star)}} p(\mathbf{z}, \mathbf{w}, \mathbf{x}; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star)) \tag{2.10}$$

3. *assignations équivalentes* : examen des assignations restantes, toutes équivalentes à $(\mathbf{z}^\star, \mathbf{w}^\star)$. Pour tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, on a
   $$\sum_{(\mathbf{z}, \mathbf{w}) \sim (\mathbf{z}^\star, \mathbf{w}^\star)} \frac{p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} = \#\,\mathrm{Sym}(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}' \sim \boldsymbol{\theta}} \frac{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}')}{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} (1 + o_P(1)) \tag{2.11}$$
   où o_P est uniforme en $\boldsymbol{\theta}$. Le maximum tient compte des configurations équivalentes alors que $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ est nécessaire lorsque $\boldsymbol{\theta}$ présente une symétrie.

Le reste de la preuve est alors directe. Puisque t_{nd} vérifie H_4, grâce à (2.9) et (2.10), le rapport de vraisemblance observée se réduit à

$$\frac{p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\theta}^\star)} = \frac{\displaystyle\sum_{(\mathbf{z}, \mathbf{w}) \sim (\mathbf{z}^\star, \mathbf{w}^\star)} p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) + p(\mathbf{x}; \mathbf{z}^\star, \mathbf{w}^\star, \boldsymbol{\theta}^\star) o_P(1)}{\displaystyle\sum_{(\mathbf{z}, \mathbf{w}) \sim (\mathbf{z}^\star, \mathbf{w}^\star)} p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}^\star) + p(\mathbf{x}; \mathbf{z}^\star, \mathbf{w}^\star, \boldsymbol{\theta}^\star) o_P(1)}$$

et le résultat final est déduit en utilisant (2.11).

2.6.3 Contrôles global et local

Les propositions suivantes énoncent les résultats importants permettant le contrôle de $\tilde{\Lambda}$ (proposition 1), et le contrôle de $F_{nd} - \tilde{\Lambda}$ global (proposition 2) et local (proposition 3).

Proposition 1 (Séparabilité pour $\tilde{\Lambda}$)**.** *Conditionnellement à Ω_1, il existe une constante $C > 0$ telle que pour tout $(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, C)$:*

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4} \left(d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}\right) \tag{2.12}$$

De plus, il existe une constante $B(C) > 0$ telle que pour tout $(\mathbf{z}, \mathbf{w}) \notin S(\mathbf{z}^\star, \mathbf{w}^\star, C)$

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -B(C)\, nd \tag{2.13}$$

Proposition 2 (larges déviations pour F_{nd})**.** *Soit* $\mathrm{Diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty$. *Pour tout* $\varepsilon_{n,d} < \kappa\bar{\sigma}$, n, d

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta},\mathbf{z},\mathbf{w}} \left\{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})\right\} \geq \bar{\sigma} nd\,\mathrm{Diam}(\boldsymbol{\Theta}) 2\sqrt{2}\varepsilon_{nd}\left[1 + \frac{gm}{2\sqrt{2nd}\varepsilon_{nd}}\right]\right) \leq g^n m^d \exp\left(-\frac{nd\varepsilon_{nd}^2}{2}\right) \quad (2.14)$$

En particulier, si n et d sont suffisamment grands pour que $2\sqrt{2nd}\varepsilon_{nd} \geq gm$, l'inégalité précédente garantit qu'avec grande probabilité, $F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})$ n'est pas plus que $\bar{\sigma} nd\ Diam(\boldsymbol{\Theta}) 4\sqrt{2}\varepsilon_{nd}$. L'inégalité de concentration utilisée dans Bickel et al. (2013) pour prouver un résultat analogue pour le SBM n'est pas suffisante ici, car elle ne peut être utilisée que pour des observations bornées, ce qui n'est évidemment pas le cas pour toutes les familles exponentielles. Nous avons développé une inégalité de type Bernstein pour les variables sous-exponentielles permettant de majorer $F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})$. Plus précisément, conditionnellement à $(\mathbf{z}^\star, \mathbf{w}^\star)$, on a :

$$\begin{aligned} F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w}) &\leq F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \mathbb{E}_{\boldsymbol{\theta}^\star}[F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})|\mathbf{z}^\star,\mathbf{w}^\star] \\ &= \sum_{kk'}\sum_{\ell\ell'} (\alpha_{k'\ell'} - \alpha^\star_{k\ell})\, W_{kk'\ell\ell'} \\ &\leq \sup_{\substack{\Gamma\in\mathbb{R}^{g^2\times m^2} \\ \|\Gamma\|_\infty \leq \mathrm{Diam}(\boldsymbol{\Theta})}} \sum_{kk'}\sum_{\ell\ell'} \Gamma_{kk'\ell\ell'} W_{kk'\ell\ell'} := Z \end{aligned}$$

uniformément en $\boldsymbol{\theta}$, où les $W_{kk'\ell\ell'}$ sont indépendants et définis par :

$$W_{kk'\ell\ell'} = \sum_i \sum_j z^\star_{ik} w^\star_{j\ell} z_{i,k'} w_{j\ell'} \left(x_{ij} - \psi'(\alpha^\star_{k\ell})\right)$$

Nous avons montré que $W_{kk'\ell\ell'}$ et Z sont des variables sous exponentielles, permettant de définir une inégalité de Bernstein sur laquelle s'appuie fortement la proposition 2.

Proposition 3 (petites déviations F_{nd})**.** *Supposons* A_4. *Conditionnellement à* Ω_1, *il existe* $C > 0$, *tel que pour tout* $\tilde{c} \leq C$,

$$\sum_{\substack{(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,\tilde{c}) \\ (\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}} \frac{\Lambda(\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z}^\star,\mathbf{w}^\star)}{d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}} = o_P(1) \quad (2.15)$$

La preuve de ce résultat utilise des propriétés de convexité des familles exponentielles et des inégalités de concentration de variables sous-exponentielles.

2.6.4 Conséquences

Les comportements asymptotiques des estimateurs du maximum de vraisemblance et variationnel se déduisent du théorème 3.

Théorème 4 (asymptotique de $\widehat{\boldsymbol{\theta}}_{MLE}$, Brault et al. (2019))**.** *Soit* $\widehat{\boldsymbol{\theta}}_{MLE}$ *l'estimateur du maximum de vraisemblance dans le modèle observé et* $\widehat{\boldsymbol{\theta}}$ *celui dans le modèle complet. Si* $\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star) = 1$, *il existe des permutations* s *de* $\{1,\ldots,g\}$ *et* t *de*$\{1,\ldots,m\}$ *telles que*

$$\hat{\boldsymbol{\pi}}(\mathbf{z}^\star) - \widehat{\boldsymbol{\pi}}^s_{MLE} = o_P\left(n^{-1/2}\right),\ \hat{\boldsymbol{\rho}}(\mathbf{w}^\star) - \widehat{\boldsymbol{\rho}}^t_{MLE} = o_P\left(d^{-1/2}\right),$$
$$\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star,\mathbf{w}^\star) - \widehat{\boldsymbol{\alpha}}^{s,t}_{MLE} = o_P\left((nd)^{-1/2}\right).$$

Si $\#\,\mathrm{Sym}(\boldsymbol{\theta}) \neq 1$, $\widehat{\boldsymbol{\theta}}_{MLE}$ *est toujours consistant : il existe des permutations* s *de* $\{1,\ldots,g\}$ *et* t *de* $\{1,\ldots,m\}$ *telles que*

$$\begin{aligned}\hat{\boldsymbol{\pi}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\pi}}_{MLE}^{s}=o_{P}\left(1\right),&\hat{\boldsymbol{\rho}}\left(\mathbf{w}^{\star}\right)-\widehat{\boldsymbol{\rho}}_{MLE}^{t}=o_{P}\left(1\right),\\ \hat{\boldsymbol{\alpha}}\left(\mathbf{z}^{\star},\mathbf{w}^{\star}\right)-\widehat{\boldsymbol{\alpha}}_{MLE}^{s,t}&=o_{P}\left(1\right).\end{aligned}$$

Ainsi, l'estimateur du maximum de vraisemblance du LBM est consistant et asymptotiquement normal, de même comportement que l'estimateur du maximum de vraisemblance du modèle complet quand $\boldsymbol{\theta}$ ne présente pas de symétrie. On montre également que ce résultat s'étend au cas de l'estimateur variationnel.

2.7 Choix de modèle

Le choix du nombre pertinent de blocs dans un LBM est évidemment d'une importance cruciale. Mais ce problème de sélection de modèle est difficile pour plusieurs raisons. D'une part, c'est un couple (g,m) de dimensions qu'il faut sélectionner au lieu d'une seule valeur. D'autre part, les critères de vraisemblance pénalisée, tels que AIC ou BIC, ne sont pas calculables directement, car le calcul de la vraisemblance maximisée n'est pas réalisable et seul un minorant est accessible. Enfin, il peut être délicat de déterminer le nombre d'unités statistiques d'un LBM (nombre de lignes, nombre de colonnes, nombre de cellules).

Cependant, il est possible de calculer explicitement la vraisemblance complète

$$p(\mathbf{x},\mathbf{z},\mathbf{w}\mid g,m)=\int_{\Theta_{g,m}}p(\mathbf{x},\mathbf{z},\mathbf{w}\mid\boldsymbol{\theta}_{g,m})p(\boldsymbol{\theta}_{g,m})d\boldsymbol{\theta}_{g,m}$$

et donc de définir un critère ICL (section 1.3.2) exact pour le LBM multinomial en utilisant les propriétés de conjugaison des lois a priori de Dirichlet définies équation 2.3, (Keribin et al., 2012; Brault, 2014; Keribin, 2015; Keribin et al., 2015) ; Ainsi,

$$\begin{aligned}\mathrm{ICL}(g,m)&=\log\Gamma(ga)+\log\Gamma(ma)-(m+g)\log\Gamma(a)\\&+mg(\log\Gamma(rb)-r\log\Gamma(b))\\&-\log\Gamma(n+ga)-\log\Gamma(d+ma)\\&+\sum_{k}\log\Gamma(z_{\cdot k}+a)+\sum_{\ell}\log\Gamma(w_{\cdot\ell}+a)\\&+\sum_{k,\ell}\Big[(\sum_{h}\log\Gamma(N_{k\ell}^{h}+b))\\&-\log\Gamma(z_{\cdot k}w_{\cdot\ell}+rb)\Big].\end{aligned}$$

où $z_{\cdot k}$, $w_{\cdot\ell}$ et $N_{k\ell}^{h}$ sont définis en (2.4). Les variables manquantes $\mathbf{z},\mathbf{w}$ ont été remplacées par leur MAP, suivant Biernacki et al. (2000) :

$$(\hat{\mathbf{z}},\hat{\mathbf{w}})=\arg\max_{(\mathbf{z},\mathbf{w})}p(\mathbf{z},\mathbf{w}|\mathbf{x};\hat{\boldsymbol{\theta}}),$$

où $\hat{\boldsymbol{\theta}}$ est l'estimation du paramètre calculée à partir de l'algorithme EM-VBayes initialisé par l'échantillonneur de Gibbs, suivant le protocole décrit dans la section 2.5.2. Une

approximation de Laplace comme dans l'équation (1.13) permet d'obtenir l'expression asymptotique suivante

$$\begin{aligned}\mathrm{ICL}(g,m) \simeq & \max_{\boldsymbol{\theta}} \log p(\mathbf{x}, \widehat{\mathbf{z}}, \widehat{\mathbf{w}}; \boldsymbol{\theta}) \\ & - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d - \frac{gm(r-1)}{2}\log(nd). \end{aligned} \tag{2.16}$$

Elle n'est d'aucune utilité pour le critère ICL en tant que tel, mais en utilisant la décomposition de McLachlan et Peel (2000) décrite à l'équation 1.15, on a :

$$\mathrm{ICL}(g,m) = \log p(\widehat{\mathbf{z}}, \widehat{\mathbf{w}} | \mathbf{x}; \widehat{\boldsymbol{\theta}}) + \mathrm{BIC}(g,m) \tag{2.17}$$

Ceci laisse conjecturer la forme d'un critère BIC

$$\begin{aligned}\mathrm{BIC}(g,m) = & \log p(\mathbf{x}; \widehat{\boldsymbol{\theta}}) \\ & - \frac{gm(r-1)+g-1}{2}\log n - \frac{gm(r-1)+m-1}{2}\log d \end{aligned} \tag{2.18}$$

De plus, sous le vrai modèle, la loi conditionnelle des labels tend vers une loi de Dirac en la vraie partition, ce qui amène à conjecturer l'équivalence asymptotique des critères BIC et ICL sous le vrai modèle, ce qui serait une propriété tout à fait remarquable du LBM.

2.8 Applications

Les travaux pendant la thèse de Vincent Brault ont été essentiellement méthodologiques et confrontés sur des données simulées. Ils ont toutefois été mis en application sur un cas d'étude de votes des députés du parlement américain [1] que nous présentons ici. Un cas d'application de pharmacovigilance été mise en oeuvre par Robert (2017) et sera détaillé au chapitre 3.

Congressional Voting Records Ce jeu de données comporte l'enregistrement des votes (données catégorielles à 3 niveaux : 'oui', 'non', 'abstention ou absent') de 435 membres du 98^{e} congrès sur 16 questions clés différentes. La méthodologie EM-VBayes initialisée par un échantillonnage de Gibbs a été exécutée avec des valeurs d'hyperparamètres $(a,b) = (4,1)$. Les critères ICL et BIC ont été calculés pour sélectionner le nombre de blocs. ICL a sélectionné le modèle $(g,m) = (5,7)$, tandis que BIC a sélectionné le modèle $(g,m) = (4,6)$. Le matrices dont les lignes et colonnes sont réorganisées suivant ces blocs sont représentés Figure 2.6 où 'oui' est en noir, 'non' en blanc, et 'abstention ou absent' en gris. Les quatre premiers clusters-colonne sont identiques dans les deux cas et permettent de bien trancher les réponses en fonction des appartenance aux partis. Les deux critères isolent dans un groupe spécifique (resp. 7 et 6) une question caractérisée par fort taux d'abstention. Le cluster-colonne 5 obtenu avec BIC est divisé en deux clusters-colonne (5 et 6) dans le modèle sélectionné avec ICL qui a raffiné la prise en compte des abstentions. Ces résultats ont été comparés à ceux de Wyse et Friel (2012) utilisant un *collapsed sampler* sur modèle binarisé ('non' et 'abstention ou absent' agrégés dans la même réponse). Cet algorithme marginalise les paramètres du modèle avec une loi a priori uniforme uniforme (i.e. $(a,b) = (1,1)$ figure 2.5). Utilisés dans des conditions similaires, les deux algorithmes donnent des résultats sensiblement identiques. L'exécution du *collapse sampler* nécessite environ six fois plus d'itérations, mais permet d'accéder à la loi a posteriori du nombre de blocs, lorsque la loi a priori est une loi de Poisson tronquée.

1. le jeu de données *Congressional Voting Records* est disponible à l'adresse http ://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records.

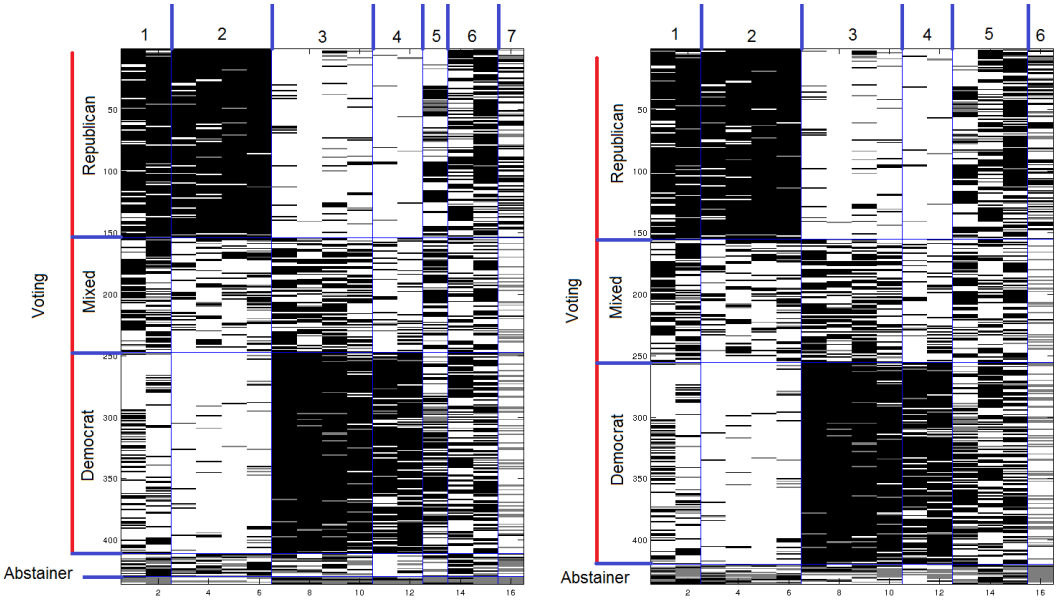


FIGURE 2.6 – Réorganisation des données suivant le co-clustering résultant du critèree ICL (à gauche) et du critère BIC (à droite)

Chapitre 3

Pharmacovigilance

Ces travaux ont été effectués pendant la thèse de Valérie Robert (Robert, 2017), en co-direction avec Gilles Celeux. Ils ont été au départ motivés par le besoin d'un outil de détection des effets secondaires médicamenteux à partir de bases de notifications en pharmacovigilance, et ont amené à entreprendre des développements méthodologiques intéressants.

Les résultats sont essentiellement méthodologiques : mise en place d'algorithmes bayésiens sur LBM avec données de Poisson ; définition du modèle des blocs latents multiples (MLBM) une extension du LBM permettant de traiter la spécificité des données individuelles de pharmacovigilance et mise en place d'une stratégie d'estimation ; développement d'une procédure gloutonne bi-km1 de parcours des modèles LBM et MLBM entre autres.

3.1 Besoin applicatif et données

Avant l'autorisation de mise sur le marché d'un médicament, des essais cliniques strictement encadrés sont réalisés sur des échantillons de petite taille et homogènes, pour une durée limitée et avec des paramètres contrôlés. C'est après l'autorisation de mise sur le marché que la veille de pharmacovigilance se met en place suivant un système de signalement spontané. Les informations sont remontées par les professionnels de santé à l'agence nationale de sécurité du médicament et des produits de santé (ANSM, antérieurement AFFSAPS) pour chaque cas individuel et peuvent être utilisées pour détecter un risque potentiel pouvant mener à générer une alerte.

L'équipe Biostatistique et Pharmacoépidémiologie [1] de Pascale Tubert-Bitter dispose d'une instance anonymisée de cette base sous forme d'une liste de médicaments et d'une liste d'effets indésirables pour chaque cas individuel d'interaction suspectée. Les données collectées entre 2000 et 2010 représentent $n = 219\ 340$ rapports individuels impliquant $J = 2142$ de médicaments et $K = 4216$ effets indésirables.

Ces données peuvent être représentées par deux matrices binaires (figure 3.1), $\mathbf{x}$ pour les médicaments, $\mathbf{y}$ pour les effets indésirables, partageant les mêmes individus en ligne :

- pour $i = 1, \ldots, n$, $j = 1, \ldots, J$, $x_{ij} = 1$ si l'individu i a pris le médicament j, 0 sinon ;
- pour $i = 1, \ldots, n$, $k = 1, \ldots, K$, $y_{ik} = 1$ si l'effet indésirable k a été observé sur l'individu i, 0 sinon.

1. INSERM UMR U1181 : Biostatistique, biomathématique, pharmacoépidémiologie et maladies infectieuses (B2PHI)

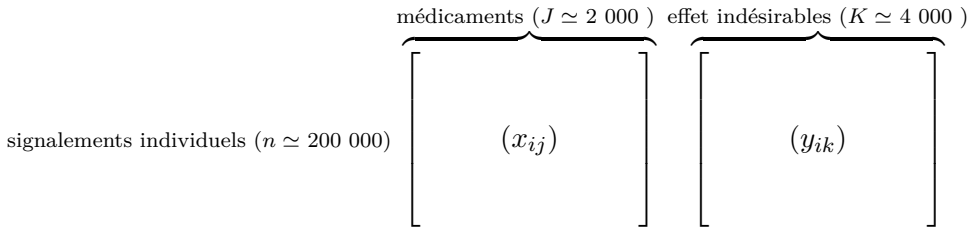


FIGURE 3.1 – Données individuelles de pharmacovigilance mises sous forme matricielle.

Les bases de pharmacovigilance sont établies par signalement spontané et peuvent donc comporter de nombreux biais : en particulier, un effet indésirable peut être plus susceptible d'être signalé sur un nouveau médicament que sur un plus ancien. Nous ne nous sommes pas focalisés sur cet aspect dans nos travaux, mais avons abordé les problèmes méthodologiques en supposant que les données à disposition ont été élaborées avec un échantillonnage statistique cohérent, ce qui pose déjà des problèmes méthodologiques intéressants.

La mise en place d'un protocole rigoureux permettant d'inclure des signalements spontanés a été proposé par Giraud et al. (2014); Coron et al. (2017) dans une application d'écologie. L'adaptation de leur méthode permettant de capitaliser des données opportunistes pour surveiller les abondances relatives des espèces à la pharmacovigilance pourrait être une étude à part entière, à mener en étroite collaboration avec l'ANSM. Leur méthode nécessite quoi qu'il en soit de savoir traiter des données correctement échantillonnées. Des jeux de référence commencent à émerger, et des signaux de type positifs et négatifs identifiés, comme par exemple dans la référence OMOP (Ryan et al., 2013).

3.2 LBM de Poisson sur données de contingence

Une des spécificités de la représentation binaire des données individuelles de pharmacovigilance est le faible taux de valeurs informatives (environ 2% de 1), ce qui rend leur traitement délicat. Ainsi, traditionnellement, c'est le tableau de contingence (médicaments, effets) $\mathbf{c}$ qui est étudié et les méthodes de détection utilisent des mesures de dissimilarité pour détecter les couples (médicament, effet) anormalement fréquents. Il est utilisé par exemple dans les méthodes *Proportional Reporting Ratio* (Evans et al., 2001), *Reporting Odds Ratio* (van Puijenbroek et al., 2002), *Gamma Poisson Shrinker* (DuMouchel, 1999).

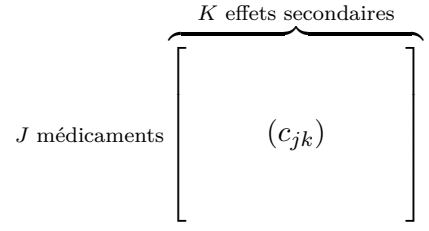


FIGURE 3.2 – Table de contingence croisant médicaments et effets secondaires : c_{ij} est le nombre de signalements pour le médicament i en conjonction avec l'effet j.

Un modèle des blocs latents peut être défini sur un tableau de contingence dont le nombre de lignes (médicaments) est I et le nombre de colonnes (effets) est J. Les blocs

sont définis par le produit cartésien d'une partition de H clusters-lignes par une partition de L clusters-colonne. Pour chaque bloc (h, ℓ), $h = 1, \ldots, H$, $\ell = 1, \ldots, L$, la densité conditionnelle $\varphi(c_{jk}, \alpha_{h\ell})$ est une densité de Poisson. Suivant Govaert et Nadif (2010), le paramètre $\alpha_{h\ell}$ fait intervenir un effet ligne μ_j et un effet colonne ν_k

$$\varphi(c_{jk}, \alpha_{h\ell}) \sim \mathcal{P}(\mu_j \nu_k \gamma_{h\ell}) \tag{3.1}$$

L'effet ligne (resp. colonne) permet de classer deux lignes (resp. colonnes) proportionnelles dans le même cluster-ligne (resp. colonne) et $\gamma_{h\ell}$ est une interaction pour le bloc (h, ℓ). Certaines conditions sont nécessaires pour assurer l'identifiabilité :

$$\sum_j \mu_j = \sum_k \nu_k = \sum_{j,k} c_{jk}.$$

Ceci assure que $\mathbb{E}(\sum_k c_{jk}) = \mu_j$ et $\mathbb{E}(\sum_j c_{jk}) = \nu_k$ et amène naturellement à estimer μ_j et ν_k au préalable par $\sum_k c_{jk}$ et $\sum_j c_{jk}$.

Nous avons développé l'inférence bayésienne du modèle LBM-Poisson (cf. section 2.5.2), en définissant les loi a priori sur les poids des clusters-médicaments $\boldsymbol{\rho}$, des clusters-effets $\boldsymbol{\tau}$, et sur les taux γ de la façon suivante :

$$\boldsymbol{\rho} \sim \mathcal{D}(a, \ldots, a), \quad \boldsymbol{\tau} \sim \mathcal{D}(a, \ldots, a), \; \gamma_{k\ell} \sim \Gamma(\alpha, \beta), \tag{3.2}$$

où $\Gamma(\alpha, \beta)$ est la loi Gamma de paramètres (α, β). Les labels des médicaments sont notés v_j, et ceux des effets w_k.

Le choix des hyper-paramètres est délicat. En particulier, il n'existe pas de choix permettant une loi non informative. Des simulations pour étudier la sensibilité des algorithmes au choix des hyperparamètres a permis de confirmer le choix régularisateur de $a = 4$ (cf. section 2.5.2). De plus, $\beta = 1$ a clairement un effet bénéfique contre la dégénérescence des classes, ainsi que tout $\beta \leq 1$. Ainsi 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 $\alpha = 1, \beta = 0.01$ (Robert, 2017, chap. 3).

Grâce au choix des loi a priori conjuguées, ICL peut être calculé explicitement, permettant l'utilisation facile de critère de choix de modèle :

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

3.3 MLBM binaire sur données individuelles

Les méthodes sur tableau de contingence ont cependant une faiblesse : par leur constitution même, elles ne permettent pas de déceler les effets de *co-prescription* (quand deux médicaments sont en général prescrits simultanément, détecter celui responsable de l'effet) ou de *masquage* (les couples fortement notifiés peuvent fausser la détection de couples plus faiblement représentés). Il est donc souhaitable de revenir aux données individuelles

pour une analyse plus pertinente, même si elles sont beaucoup plus volumineuses. Ceci motive certaines études utilisant des outils de régression : régression logistique avec étape de présélection (Harpaz et al., 2013) ou parcimonieuse (Marbac et al., 2016) par exemple. Mais celles-ci ne peuvent prendre en compte plusieurs effets de façon simultanée.

Nous avons étendu le LBM à l'analyse de données individuelles décrites par plusieurs ensembles de variables en définissant le *modèle à blocs latents multiples* (MLBM), un modèle de co-clustering qui permet de regrouper simultanément plusieurs ensembles distincts de variables observées sur les mêmes individus. Ainsi, ce modèle peut être vu comme un LBM sous contrainte de structure des clusters-colonne. Les hypothèses adaptent celles du LBM

$\mathbf{H}_1'$: produit cartésien *Les blocs sont le résultat du produit cartésien d'une partition* $\mathbf{z}$ *des lignes en* G *clusters-ligne par l'union d'une partition* $\mathbf{v}$ *des colonnes-médicaments en* H *clusters-médicaments et d'une partition* $\mathbf{w}$ *des colonnes-effets en* L *clusters-effets*

$\mathbf{H}_2'$: indépendance des labels *Les variables latentes* $\mathbf{z}$, $\mathbf{v}$ *et* $\mathbf{w}$ *sont indépendantes et indépendantes entre elles.*

$\mathbf{H}_3'$: indépendance conditionnelle *Les variables* $\mathbf{x} = (x_{ij})$ *et* $\mathbf{y} = (y_{ik})$ *sont indépendantes conditionnellement aux labels* $(\mathbf{z}, \mathbf{v}, \mathbf{w})$ *de la façon suivante :*

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}, \mathbf{v}, \mathbf{w}) = p(\mathbf{x}|\mathbf{z}, \mathbf{v})p((\mathbf{y}|\mathbf{z}, \mathbf{v})$$

La loi conditionnelle de x_{ij} *(resp.* y_{ik}*) dépend uniquement de son bloc d'appartenance* (g, ℓ) *(resp.* (g, h)*).*

Les algorithmes d'estimation (VEM, EM-VBayes et Gibbs) ont été développés (graphe du modèle bayésien représenté sur la figure 3.3). Une étude de simulation a été conduite pour étudier la sensibilité aux hyperparamètres, en particulier dans le cas de matrices très creuses, qui est celui la pharmacovigilance. Le choix $a = 4$ permettant de limiter la dégénérescence des classes n'est pas remis en cause ; en revanche, choisir $b = 1$ comme préconisé en 2.5.2 revient à considérer la loi des paramètres α des Bernoulli comme uniforme, alors qu'ils sont en général (très) petits pour les matrices creuses. L'échantillonneur de Gibbs y est particulièrement sensible, et on peut penser raffiner la définition de la loi a priori $\alpha \sim \mathcal{B}e(b_1, b_2)$ en fonction de la densité des observations : par exemple, utiliser une loi $\mathcal{B}e(2, 100)$ pour des paramètres α de l'ordre de 2%. Bien que cette loi impose un a priori très marqué, elle a donné des résultats satisfaisants dans des expériences préliminaires. L'algorithme EM-VBayes semble moins sensible, et peut continuer à être utilisé avec les valeurs de référence $(a, b) = (4, 1)$.

Note Si la loi conditionnelle est binaire pour chacun des deux tableaux $\mathbf{x}$ et $\mathbf{y}$ de données individuelles en pharmacovigilance, cette méthodologie s'applique bien sûr au cas où elles seraient différentes. De même, l'extension à plus de deux matrices est naturelle (Jacques et Biernacki, 2018).

3.4 Procédure gloutonne de parcours des modèles

La sélection de modèle d'un LBM nécessite de comparer les valeurs du critère ICL pour tous les couples (G, H) définissant le nombre de clusters-ligne et clusters-colonne. Dans un MLBM, il faut théoriquement prendre en compte toutes les combinaisons des

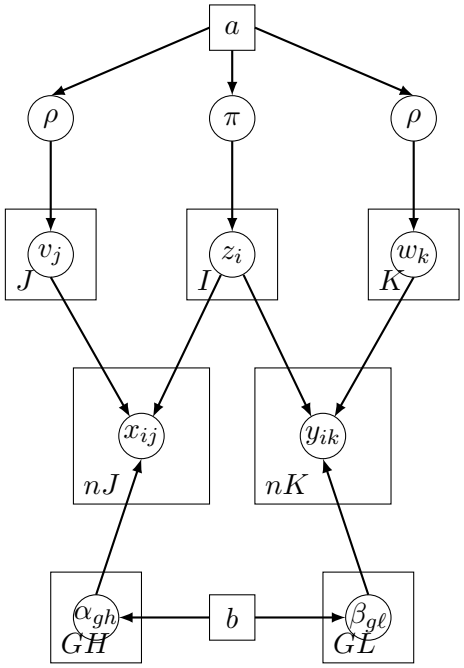


FIGURE 3.3 – Graphe bayésien du modèle MLBM.

triplets (G, H, L) (clusters-individus, clusters-médicaments, clusters-effets). Ces explorations coûtent cher dès que la dimension des modèles augmente. Dans Robert et al. (2016); Robert (2017) est proposé un algorithme glouton appelé Bi-KM1, dont le principe pour le LBM est le suivant : à partir d'un co-clustering à (G, H) blocs, tous les modèles à $(G+1, H)$ blocs sont analysés. Le calcul sur chacun d'eux est initialisé par une division aléatoire en deux de l'un des G clusters-ligne initiaux, et répété pour chacun de ces clusters. Le même principe est utilisé pour calculer l'ICL sur un modèle à $(G, H + 1)$ blocs. Le modèle avec la meilleure valeur de ICL est sélectionné.

Bien que cet algorithme de recherche soit sous-optimal, il bénéficie d'une stratégie d'initialisation bien plus efficace qu'une initialisation aléatoire sur chaque modèle. De plus, il permet un parcours glouton de l'espace des modèles (*greedy procedure*), évitant de parcourir $G_{max} \times H_{max}$ modèles. Le principe est identique pour le cas du MLBM, avec trois directions de recherche, comme représenté schématiquement figure 3.4.

3.5 Méthodologie en pharmacovigilance

Ni l'utilisation de méthodes explicatives standard sur le jeu de données de l'ANSM (section 3.1), ni le LBM Poisson (section 3.2) ne permettent de détecter un des couples de référence (médicaments, effets) déjà rapporté par l'OMOP. Ceci pourrait s'expliquer par le fait que l'ensemble des couples références n'est pas exhaustif, par la faiblesse du signal dans la base de données française ou par un artefact de co-prescription.

Afin de renforcer le signal, nous avons construit un tableau de contingence réduit aux seuls individus ne prenant qu'un seul médicament et n'ayant qu'un effet, abandonnant l'artefact de co-prescription : cela représente environ 20% des rapports et concerne 1482 de médicaments et 2239 effets indésirables. ICL sur LBM Poisson sélectionne à modèle à 19×20 blocs. Les signaux détectés par LBM dans le tableau de contingence complet sont regroupés dans des blocs du tableau de contingence réduit de paramètres d'intensité

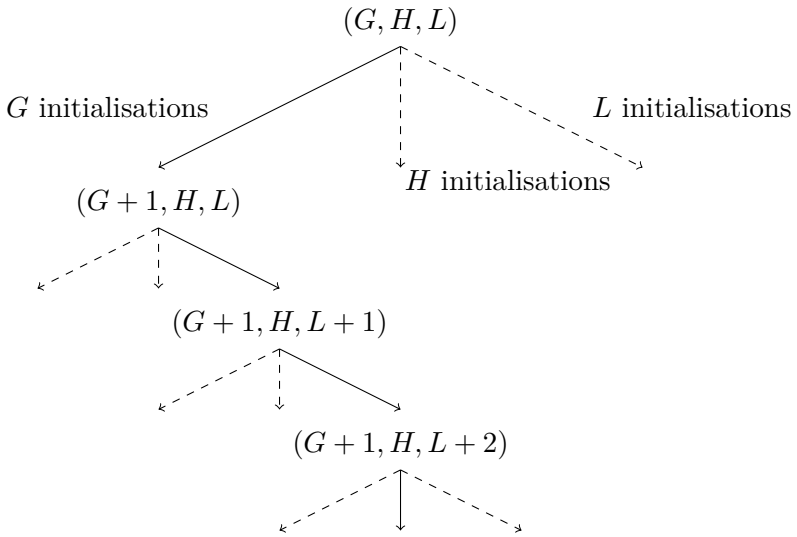


FIGURE 3.4 – Représentation schématique de l'algorithme Bi-KM1 pour le MLBM

les plus élevés. Cependant, les signaux de référence OMOP ne sont généralement pas regroupés dans les mêmes blocs.

Plusieurs difficultés se posent lors du traitement des données individuelles : d'une part, les données sont maintenant clairsemées (environ 2% de 1) et les hyperparamètres doivent être soigneusement étalonnés comme indiqué à la section 3.3. D'autre part, les performances sont affectées par la taille des données et les trois dimensions du modèle. Pour pallier ce dernier problème, nous proposons d'initialiser les groupes de médicaments et d'effets du MLBM avec ceux obtenus à partir du tableau de contingence. Cette stratégie donne des résultats encourageants sur des données simulées : sur le tableaux de contingence, l'algorithme ayant tendance à regrouper les signaux des blocs d'intensité plus élevée. Sur les données individuelles, le modèle sélectionné permet en général de concentrer les signaux sur quelques blocs d'intensité élevée (valeurs élevées pour α_{gh} et $\beta_{g\ell}$) ne possédant en général que peu de médicaments. De plus, le processus a tendance à séparer les individus qui ont pris un médicament avec effet de ceux qui les ont pris sans effet.

Si ces résultats sont encourageants, la procédure n'a cependant pas pu être mise en œuvre sur la totalité de la base de l'ANSM à cause de contingences informatiques (taille des objets manipulés) et de temps pour les résoudre. Ce point reste une problématique intéressante à étudier.

Chapitre 4

Neuro-imagerie

Ces travaux ont été effectués dans le cadre de la thèse de Vincent Michel (2010), *Understanding the visual cortex by using classification techniques*, en co-direction avec G. Celeux et B. Thirion. Ils proposent deux développements méthodologiques permettant de traiter des données en grande dimension dans un cadre supervisé que sont les cartes d'activation neuronales : (1) effectuer une régularisation bayésienne grâce à un a priori sous forme modèle de mélange sur les voxels ; (2) Élaguer un arbre de classification (non supervisée) hiérarchique sur les voxels en utilisant un critère prédictif, d'où le nom *clustering supervisé* qui peut paraître un oxymore.

Le *codage neuronal* est la correspondance entre un stimulus (la vision d'une forme, d'une taille, d'un nombre par exemple) et sa représentation par un individu activant un ensemble de réponses neuronales. Accéder à ce codage neuronal peut être utile pour comprendre les processus mentaux, et plus généralement, la façon dont le cerveau traite l'information. L'étude de ce codage neuronal peut être réalisée à différentes échelles ou structures (du neurone simple à une grande population de neurones telle que les colonnes corticales avec environ 10^4 neurones), appelées *entités codantes*. De plus, cette étude peut être réalisée par *décodage*, c'est-à-dire par reconstruction du stimulus ou de certains de ses aspects, à partir du signal qu'il provoque dans les différentes entités codantes.

Dans ces travaux, nous nous sommes concentrés sur la notion de *substrat* du codage neuronal, i.e. le nombre d'entités codantes et leur organisation spatiale. En particulier, nous avons traité deux questions fondamentales :

1. la sélection des populations neuronales impliquées dans une tâche cognitive. Si quelques neurones concentrent l'information on parle de codage *sparse* ; tandis que le codage en *population* s'observe pour de grandes populations de neurones faiblement sélectifs, ne répondant pas uniquement à un seul stimulus. Dans ce dernier cas, chaque concept est représenté par de nombreux neurones, et chaque neurone prend part à la représentation de plusieurs concepts. Ainsi, c'est tout le motif (*pattern*) d'activation qui doit être décodé.

2. la distribution spatiale de ces populations de neurones au sein du cerveau entier (cortex cérébral, ganglions de la base, thalamus) : elles peuvent être regroupées spatialement (*clustered coding*) ou largement disséminées (*distributed coding*).

Les données se présentent sous la forme de signaux enregistrés sur des petits volumes reconstituant le cerveau et appelés *voxels*. Une observation contient de l'ordre de 10^5 *voxels*, très supérieur à la taille de l'échantillon d'environ une centaine d'images : ce sont des données de grande dimension.

4.1 IRM fonctionnelle

La neuro-imagerie fonctionnelle (imagerie cérébrale fonctionnelle) vise à révéler les activités physiologiques du cerveau et leur distribution spatiale, permettant ainsi d'étudier le substrat du codage neuronal. L'IRM fonctionnelle - IRMf ou *fMRI* en anglais - est une méthode largement utilisée parmi les méthodes d'imagerie cérébrale fonctionnelle. Elle est non invasive, présente une bonne résolution spatiale (1mm), couvre la totalité du cerveau, et mesure l'intensité du signal BOLD (*blood oxygenation level-dependent*). Ce signal dépend du taux d'oxygénation du sang s'écoulant dans les petites veines du cerveau, et donne un accès indirect à l'activité neuronale : des neurones impliqués dans une tâche demandent un apport d'oxygène augmentant le signal BOLD autour de la région considérée. Cependant, l'effet reposant sur une voie métabolique est complexe, et le signal IRMf peut refléter une activité loin des neurones activés. Cette méthode est connue pour avoir une bonne résolution spatiale et malgré ses limitations, les données IRMf peuvent être utilisées pour des inférences précises et en particulier pour étudier le codage neuronal.

Les méthodes de décodage sont des méthodes supervisées dans un cadre de données de grande dimension. Nous verrons dans ce chapitre l'apport des méthodes non supervisées à ce cadre.

4.2 Modélisation

Au cours d'une expérience d'IRMf, des balayages successifs permettent l'acquisition du signal sur les *voxels*. Certains pré-traitements doivent être effectués pour extraire les informations pertinentes. Il faut par exemple corriger le décalage de temporalité de la mesure aux différents voxels dû au processus de balayage, effectuer un réalignement spatial pour corriger le mouvements potentiels apparaissant lors de l'acquisition (mouvement de la tête, dérive spatiale due au réchauffement du scanner) et appliquer un lissage spatial pour limiter les bruits haute fréquence de l'acquisition.

Inférence classique L'inférence classique repose sur la définition d'un modèle linéaire (Friston et al., 1994).

Un scan produit la mesure du signal en p voxels, représenté comme les p composantes d'un vecteur appelé image. Soit $Y \in \mathbb{R}^{n \times p}$ la matrice des n scans de p voxels et soit $X \in \mathbb{R}^{n \times r}$ la matrice du plan d'expérience codant les variables de l'expérience (facteurs expérimentaux, modélisation de nuisance, dérive basse fréquence par exemple). Le modèle linéaire s'écrit $Y = X\beta + E$ où la matrice E représente le bruit résiduel modélisé par un processus auto-régressif gaussien et où les lignes de β peuvent elles aussi être vues comme des images, appelées *cartes d'activation*. Les tests de significativité d'un effet d'intérêt sont effectués voxel par voxel de façon indépendante, conduisant à une carte de significativité appelée *Statistical Parametric Map*. Cependant, malgré la simplicité et la précision des cartes obtenues, l'inférence classique souffre d'un inconvénient majeur : l'analyse faite en chaque voxel séparément ne peut exploiter les corrélations existant entre les différentes régions du cerveau pour améliorer l'inférence. De plus, la puissance statistique est limitée par le problème de la comparaison multiple.

Par ailleurs, l'analyse multivariée classique de type Manova ou Mancova ne peut être effectuée que pour un faible nombre de voxels (et inférieur au nombre de scans) sous peine de dégénérescence.

Inférence inverse Afin de faire face à ces limitations, une approche basée sur des méthodes d'apprentissage automatique a été introduite (Dehaene et al., 1998). Il s'agit d'évaluer la spécificité de plusieurs régions du cerveau pour certaines fonctions cognitives ou perceptives, en évaluant la précision de la prédiction d'une variable comportementale d'intérêt y - la *cible* - basée sur la carte d'activation mesurée dans ces régions. Bien que notée β dans la section précédente, ces cartes seront désignées par x dans la suite, pour suivre les notations conventionnelles. Après l'étape d'acquisition, n cartes d'activation sont générées, puis utilisées pour déchiffrer le codage neuronal. Cette reconnaissance de motif (*pattern recognition*) s'appuie sur les points suivants :

- Définition de la fonction de prédiction $y = f(x, w, b)$, capable de prédire la cible y pour une nouvelle image IRMf x. Le paramètre w représente le poids de chaque *feature* définissant le pattern du substrat neuronal. Le paramètre b est un coefficient d'intercept. La cible y peut appartenir à $\mathbb{R}$ en régression ou à $\{1, \ldots, K\}$ en classification supervisée.
- Réduction de dimension pour éviter le surajustement dû à la grande dimension
- Calcul de la performance de généralisation pour appréhender la précision prédictive du modèle sur de nouvelles images. Ce sont essentiellement les méthodes de validation croisée *Leave One Out* (ou *Leave One Subjet Out* pour l'étude inter-sujet) qui sont utilisées.

De nombreuses méthodes d'apprentissage ont été utilisées pour l'inférence de la fonction de prédiction dans le cadre de la neuro-imagerie : SVM (*Support Vector Machine*, Cox et Savoy (2003)), modèles génératifs (Bayes naïf, analyse discriminante linéaire ou quadratique avec régularisation de la matrice de variance (Varoquaux et al., 2010)), méthodes de régularisation du modèle linéaire (ridge, lasso, elastic net (Carroll et al., 2009; Ryali et al., 2010)) ou régularisation bayésienne (Friston et al., 2008).

La réduction de dimension permet de supprimer des variables (*features*) non pertinentes : le modèle prédictif n'est défini que sur celles qui le sont. Celle-ci peut être native à certaines méthodes précédentes (par exemple la régression lasso), ou être opérée par des méthodes dédiées. En IRMf, la définition de région d'intérêt (*ROI*) peut utiliser des caractéristiques anatomiques a priori ou résulter de la sélection univariée de *features* basée suivant le seuillage d'un score $g(x_j)$. En revanche, la sélection à partir d'un score multivarié $g(x_{j_1}, \ldots, x_{j_k})$ souffre d'une complexité combinatoire qui en fait une méthode peu utilisée, sauf à la marier avec une procédure de type *forward*, *backward* ou *stepwise* (Michel et al., 2008).

Exigences d'un modèle en IRMf Les méthodes d'apprentissage automatique précédentes sont en général utilisées sans tenir compte des particularités des données IRMf comme la structure spatiale de l'image. De plus, les modèles obtenus sont souvent difficiles à interpréter : les cartes pondérées ne permettent pas une interprétation claire du substrat du codage neuronal, en comparaison des cartes SPM. Enfin, la réduction de dimension n'est souvent utilisée que pour augmenter la précision de la prédiction sans viser à créer des cartes utiles pour les études neuro-scientifiques. Pour tenir compte de ces considérations, un bon algorithme d'apprentissage en IRMf doit (i) être basé sur un modèle multivarié pour prendre en compte la contribution du signal sur l'ensemble des voxels (ii) prendre en compte la structure spatiale des données et leur redondance locale (iii) travailler à multi-échelle pour optimiser la définition des régions prédictives.

Deux contributions majeures réalisées pendant la thèse de Vincent Michel proposent de répondre à ces exigences et seront détaillées dans les sections suivantes :

1. une approche bayésienne multi-classes pour opérer une régularisation parcimonieuse (Michel et al., 2010b, 2009, 2011)
2. une méthode de clustering supervisé prenant en compte de l'information spatiale contenue dans les images fonctionnelles (Michel et al., 2010a, 2012)

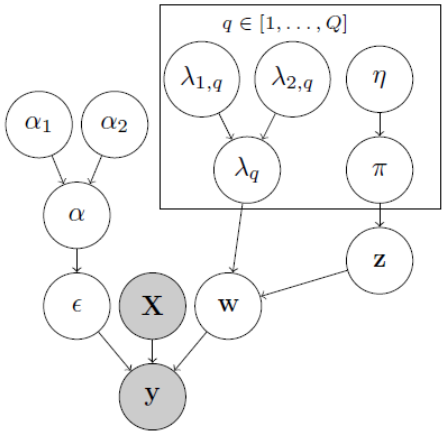
4.3 Méthode bayésienne multi-classes parcimonieuse

Contrairement aux alternatives classiques précédemment décrites, la méthode *MSBR* (*Multiclass Sparse Bayesian Regression*) regroupe les *features* dans plusieurs classes, et adapte automatiquement le niveau de régularisation de chaque classe aux données disponibles. Nous verrons que cette méthode peut être vue comme un intermédiaire entre la régression ridge bayésienne (Bishop (2006)) et l'ARD (*Automatic Relevance Determination*,Tipping (2000)), en réduisant le nombre de paramètres estimés par l'ARD tout en étant beaucoup plus adaptative que la régression bayésienne. Un autre grand atout de la méthode MSBR pour l'inférence inverse en IRMf est la création d'un regroupement de *features*, produisant ainsi des cartographies utiles du cerveau.

Modèle Nous nous plaçons dans le cadre d'un stimulus y scalaire (par exemple la taille, la position 1-d ou la cardinalité) et d'une fonction de régression f linéaire. Pour n observations, le modèle s'écrit matriciellement

$$\mathrm{y} = \mathrm{X}\mathrm{w} + b\mathbb{1} + \epsilon \tag{4.1}$$

où les poids des voxels w et l'intercept b sont les paramètres à estimer. Rappelons que $\mathrm{w} \in \mathbb{R}^p$ peut être vu comme une image. Comme n est de l'ordre de 100, et p de l'ordre de 10^5, l'estimation de w est mal posée et nous avons utilisé l'inférence bayésienne pour ses propriétés régularisatrices. Le modèle génératif est défini et représenté sur la figure 4.1.



$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\,\mathbf{w} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \alpha^{-1}\mathbf{I_n}) \\
\alpha &\sim \Gamma(\alpha_1, \alpha_2) \\
\mathbf{z} &\sim \prod_{j=1}^{j=p}\prod_{q=1}^{q=Q} \pi_q^{\delta_{jq}} \\
\pi &\sim Dir(\eta) \\
\mathbf{w} &\sim \mathcal{N}(0, \mathbf{A}^{-1}) \\
&\text{with} \quad diag\,(\mathbf{A}) = \{\lambda_{z_1}, ..., \lambda_{z_p}\} \\
\lambda_q &\sim \Gamma(\lambda_{1,q}, \lambda_{2,q})
\end{aligned}$$

FIGURE 4.1 – Modèle graphique de la régression bayésienne multi-classes parcimonieuse (MCBR)

— La loi a priori sur le bruit ϵ est classique en régression
— Pour combiner la parcimonie de l'ARD avec la stabilité de la régression ridge bayésienne, nous avons introduit une représentation intermédiaire, sous la forme d'une partition des *features* en Q classes, chaque *feature* $j = 1, \ldots, p$ appartenant à une

classe z_j parmi les Q. Ainsi, la loi a priori des poids w est un modèle de mélange gaussien

$$f_{\mathrm{w}}(\mathrm{w}) = \prod_j f_w(w_j) = \prod_j \sum_q \pi_q f_{\mathcal{N}}(w_j; 0; \lambda_q^{-1})$$

— Tous les *features* d'une même classe $q = 1, \ldots, Q$ partagent le même paramètre de précision λ_q pour la définition de la loi a priori sur les poids w, qui par ailleurs est classique en régression ridge bayésienne.

— Les lois a priori sur α et λ définissent un modèle complètement bayésien et évitent de générer des modèles dégénérés.

Ce schéma permet de passer de la régression ridge bayésienne quand $Q = 1$, i.e. $\lambda_{z_1} = \ldots = \lambda_{z_p}$ à l'ARD quand $Q = p$, $z_j = j$ pour tout $j = 1, \ldots, p$, et tous les λ_j sont différents. L'ARD traite la parcimonie dans l'espace des hyperparamètres, contrairement à la régression lasso qui la traite dans l'espace des *features*.

Inférence Le modèle comportant des données latentes z, son ajustement peut se calculer par un algorithme EM, nécessitant le calcul de la loi conditionnelle $p(\mathrm{z}|\mathrm{X}, \mathrm{y})$ des données latentes. Or celle-ci n'est pas explicite ici. Comme déjà vu, il est alors possible d'utiliser un algorithme bayésien variationel ou un échantillonneur de Gibbs.

Le modèle MSBR nécessite peu d'hyperparamètres ; nous les avons choisis légèrement informatifs et spécifiques à la classe afin de refléter un large éventail de comportements possibles pour la répartition des poids. Ce choix est équivalent à définir des distributions de Student à queues lourdes, avec des variances à différentes échelles, comme a priori sur le paramètres de poids.

Simulations Des expérimentations ont été faites sur des données simulées avec ces deux algorithmes comparées à l'état de l'art de méthodes de régularisation (elastic net, SVR, régression ridge bayésienne, ARD).

L'échantillonneur de Gibbs sur le modèle MSBR surpasse les autres méthodes classiques testées, et apporte une précision de prédiction plus stable. De plus MSBR crée un clustering de *features* basé sur leur pertinence, et extrait explicitement des groupes de *features* informatifs. Enfin, en adaptant la régularisation des différents groupes de voxels, MSBR arrive à retrouver le vrai support des poids, et une représentation parcimonieuse. Ces propriétés se retrouvent lors de l'étude d'une expérimentation de représentation mentale de la taille.

Dans certaines expériences et sans surprise, l'algorithme bayésien variationel donne des prévisions moins précises que celles obtenues par l'échantillonnage de Gibbs, ce qui peut s'expliquer par la difficulté d'initialiser les différentes variables (en particulier z), l'amenant souvent à être piégé dans des minima locaux.

Conclusion La modèle MSBR généralise les approches classiques d'ARD et de régression ridge bayésienne pour l'étude de cartes d'activation IRMf.

En effectuant une régularisation différente pour les *features* pertinents et non pertinents, il permet à la fois une sélection d'un sous-ensemble de *features* et une estimation précise des poids du modèle. Il peut s'adapter facilement, dans un cadre bayésien, aux différents niveaux de parcimonie des données IRMf. La régularisation est adaptative comme en ARD, mais elle est effectuée avec beaucoup moins d'hyper-paramètres, et est donc moins sujette au surajustement dans l'espace des hyper-paramètres.

La question de la sélection du modèle (c'est-à-dire le nombre de classes Q) n'a pas été abordée. L'utilisation de l'énergie libre dans un critère pénalisé ne semble pas une

approche prometteuse en raison de l'instabilité de l'algorithme variationel sur MSBR. Une méthode plus intéressante est celle détaillée dans Chib et Jeliazkov (2001), qui peut être utilisée dans l'algorithme d'échantillonnage de Gibbs. Ici, la sélection du modèle est effectuée implicitement en vidant les classes qui ne correspondent pas bien aux données. À cet égard, le choix de lois a priori hétérogènes pour différentes classes est crucial.

4.4 Clustering supervisé

Les méthodes précédentes ne prennent pas en compte la contrainte spatiale des voxels, les évaluant souvent de manière indépendante : les voxels peuvent être redondants ou disséminés dans de larges zones du cerveau, ce qui rend leur interprétation délicate. Assurer une cohérence spatiale doit permettre la construction de régions informatives et anatomiquement cohérentes.

Définir une information spatiale Une première solution consiste à introduire l'information spatiale sur les voxels, par exemple en ajoutant des lois a priori basées sur les régions (Palatucci et Mitchell, 2007), ou en utilisant une régularisation sur l'espace (Michel et al., 2008).

L'agglomération de *features* est un moyen naturel d'utiliser l'information spatiale, consistant à remplacer le signal acquis voxel par voxel, par des moyennes locales calculées sur des *parcelles* (Flandin et al., 2002; Thirion et al., 2006), formant une partition de l'ensemble des voxels. Le nombre de *features* est ainsi réduit d'environ 10^5 voxels à 10^2 parcelles. De plus les moyennes sont robustes pour les petits déplacements spatiaux et moins sujettes à la variabilité inter-sujet que les méthodes basées sur voxels. Le partitionnement peut être purement géométrique, mais pour prendre en compte simultanément l'information spatiale et l'information fonctionnelle, des méthodes de clustering ont aussi été proposées : clustering spectral (Thirion et al., 2006), mélanges gaussiens (Thyreau et al., 2006), K-moyennes (Ghebreab et al., 2008) par exemple. Cependant, le nombre de clusters peut être difficile à établir, et les moyennes locales peuvent faire perdre une information finement localisée.

La méthode *searchlight* (Kriegeskorte et al., 2006) ne conserve que les voxels voisins pour le modèle prédictif, et est donc largement utilisée pour l'étude locale. Mais elle ne peut pas gérer les interactions à longue portée dans le codage de l'information et souffre du problème des comparaisons multiples.

Clustering supervisé Nous avons proposé une méthode permettant de s'affranchir des inconvénients des méthodes précédentes en effectuant un regroupement supervisé des *features* par agglomération plutôt que par simple sélection. Cet algorithme permet de prédire l'état cognitif cible d'un sujet en sélectionnant simultanément les régions connexes du cerveau active pendant cet état. La prédiction de la cible est utilisée dès l'étape de clustering pour produire une segmentation adaptative de *features* en régions étendues ou plus finement localisées suivant la force du signal et qui peut donc être considérée comme multi-échelle.

Le principe de la méthode est représenté figure 4.2

Dans une première étape *bottom-up*, un clustering hiérachique ascendant est effectué sur les voxels par la méthode de Ward, avec contrainte de n'autoriser l'agrégation de clusters que s'ils sont adjacents spatialement. La méthode produit des clusters spatialement connectés, c'est-à-dire des parcelles. La succession de parcellisations emboîtées est repré-

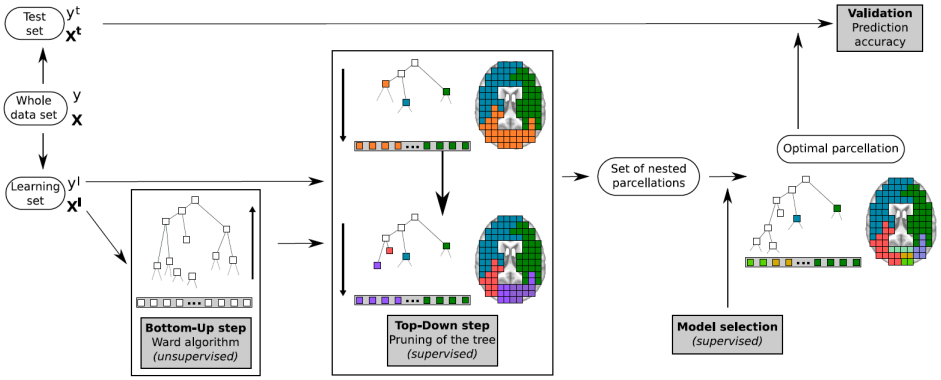


FIGURE 4.2 – Schéma de l'algorithme de clustering supervisé

sentée par un arbre binaire $\mathcal{T}$ et n'importe quelle découpe de l'arbre peut être identifiée à une parcellisation

Dans une deuxième étape *top-down*, l'arbre est élagué pour définir des parcellisations emboîtées ayant de 1 à Δ parcelles $\mathcal{P}_1, \ldots, \mathcal{P}_\Delta$. L'élagage suivant le critère de Ward ayant servi à construire la classification hiérarchique est horizontal. Cet élagage est non supervisé puisqu'il ne fait pas appel à l'utilisation de la cible y. Nous avons proposé au contraire d'utiliser un critère prédictif pour diriger l'élagage au sein d'une stratégie gloutonne. Á chaque étape, la parcelle découpée est celle qui permet d'optimiser le score de prédiction d'un algorithme prédictif $\mathcal{A}$ parmi l'ensemble des parcelles candidates à la découpe. Cette méthode supervisée permet une plus grande flexibilité de la forme des parcellisations, et son adaptation à l'objectif de prédiction. La figure 4.3 illustre la différence entre les deux approches.

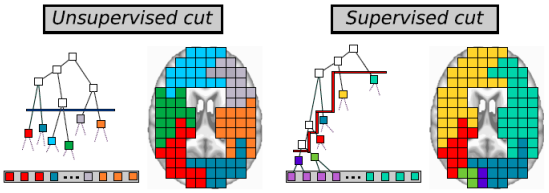


FIGURE 4.3 – Les deux méthodes d'élagage : non supervisée avec le critère de Ward (à gauche) avec un élagage horizontal, supervisée suivant un score prédictif (à droite) avec un élagage piloté par un critère prédictif

La troisième étape est celle du *choix* de la parcellisation $\widehat{\mathcal{T}}$ optimisant le critère prédictif parmi l'ensemble emboîté résultant de l'étape *top-down*. La parcellisation correspondante est alors utilisée pour calculer l'erreur de généralisation sur un échantillon test.

Dans le cas de la régression, le critère prédictif que nous avons utilisé est le ratio des variances expliquées ζ

$$\zeta(\mathrm{y}^v, \widehat{\mathrm{y}}^v) = \frac{\mathrm{var}(\mathrm{y}^v) - \mathrm{var}(\mathrm{y}^v - \widehat{\mathrm{y}}^v)}{\mathrm{var}(\mathrm{y}^v)}$$

où y^v sont les valeurs de la cible dans l'échantillon test et $\widehat{\mathrm{y}}^v$ leur prédiction à partir de l'estimation de $\mathcal{A}$ sur l'échantillon d'apprentissage, tandis qu'en classification c'est le score

classique de l'erreur de classification

$$err(\mathrm{y}^v, \widehat{\mathrm{y}}^v) = \frac{\sum_{i=1}^{n^v} \mathbb{1}\{\mathrm{y}^v \neq \mathrm{y}^v\}}{n^v}$$

où n^v désigne la taille de l'échantillon de validation.

Considérations algorithmiques Les étapes d'élagage de l'arbre et de sélection du modèle sont incluses dans une procédure de validation croisée interne au sein du jeu d'apprentissage. Cependant, ce schéma interne de validation croisée soulève différents problèmes. Premièrement, il est très fastidieux d'inclure les deux étapes dans une validation croisée interne complète. Un deuxième problème, plus crucial, est que l'exécution d'une validation croisée interne sur les deux étapes donne autant de sous-arbres optimaux que le nombre de sous-ensembles définis pour effectuer la validation croisée. Or il n'est pas facile de combiner ces différents sous-arbres optimaux afin d'obtenir un sous-arbre moyen pouvant être utilisé pour la prédiction sur l'ensemble de test. De plus, les différents sous-arbres optimaux ne sont pas construits en utilisant tout le jeu d'apprentissage et peuvent donc être soumis à un choix spécifique de la validation croisée interne. Nous avons donc choisi une heuristique empirique, potentiellement biaisée, qui consiste à utiliser deux schémas de validation croisée distincts pour l'élagage de l'arbre et l'étape de sélection du modèle.

Grâce à la réduction drastique du nombre *features* de cette approche (en général, $\Delta << p$), l'algorithme peut être facilement utilisé pour chercher des régions informatives en très grande dimension, même si les méthodes $\mathcal{A}$ elles mêmes n'y sont pas adaptées, alors que d'autres méthodes ne passent pas bien à l'échelle. Mais le bénéfice de la parcellisation a un coût, avec une augmentation du temps de calcul due à la construction de l'arbre et son élagage, restant cependant tout à fait abordable pour les analyses de données de neuro-imagerie standard.

Résultats Le clustering supervisé a été mis en œuvre sur des données simulées et des données réelles, et comparé avec un clustering non supervisé et des algorithmes de référence (*elastic net* et SVR).

En termes de précision de la prédiction, notre méthodologie donne de meilleurs résultats pour l'étude inter-sujets. L'élagage supervisé donne une précision de prédiction similaire ou supérieure à celle de l'élagage non supervisé, en particulier dans le cas où l'information est finement localisée (représentation mentale de la forme par exemple). Dans les cas où l'information est plus diffuse (représentation mentale de la taille par exemple) les méthodes non supervisées donnent de bonnes prédictions, et la différence avec notre méthode n'est pas significative.

En termes d'interprétabilité, les résultats sur les simulations et les données réelles montrent que cette approche est particulièrement apte à mettre en évidence les régions d'intérêt, tout en laissant les régions non informatives non segmentées : c'est bien une approche à plusieurs échelles. Le clustering supervisé permet de localiser des régions prédictives contiguës et créer des cartes interprétables, et peut donc être considéré comme un intermédiaire entre la cartographie du cerveau et l'inférence inverse.

4.5 Boucler la boucle ?

Que ce soit par la définition de lois a priori de type mélange sur la partition des voxels (pour MCBR) ou sur l'utilisation d'une classification non supervisée de variables guidée

par un critère de prédiction sur des individus (pour le clustering supervisé), il s'agit de définir des moyens de régulariser pour traiter le problème de données en (très) grande dimension tout en conservant une interprétation possible pour des applications.

Le co-clustering est un outil particulièrement efficace pour définir simultanément des clusters d'individus et des clusters de variables. Récemment, Tokuda et al. (2014) a utilisé un modèle probabiliste de co-clustering dans le cas d'étude d'IRMf de sujets au repos, leur état se différentiant par une condition de traitement (ici, contre la dépression). L'utilisation d'un modèle des blocs latents avec une loi conditionnelle spécifique a permis de mettre en lumière des clusters de variables corrélées, c'est-à-dire une parcellisation. L'information est résumée pour chaque individu par la valeur moyenne du signal sur chaque parcelle. L'étude de la corrélation entre les valeurs moyennées d'une parcelle et le score HRSD (caractérisant la gravité de la dépression) a montré des parcelles significatives, identifiant des zones cérébrales potentiellement impliquées dans la dépression.

Si cette méthode a permis de définir de parcellisation, elle ne garantit pas, malgré la loi conditionnelle spécifique permettant une corrélation des variables dans un bloc, que les parcelles soient connexes. Ajouter cette contrainte dans le modèle des blocs latents pourrait être une étude intéressante.

Chapitre 5

La section des curiosités

Un chercheur se doit d'être curieux, en s'ouvrant à d'autres domaines. Même si les recherches n'aboutissent pas, elles sont enrichissantes. L'étude de cas réels, indépendamment de leur utilité pratique pour le domaine concerné, est source de questions théoriques et méthodologiques intéressantes. Je présente ici trois exemples d'applications qui m'ont permis d'élargir mon domaine de recherche.

5.1 Phylogénie

Cette section présente un travail ancien mené en collaboration avec Marie-Anne Poursat (LMO), qui a donné lieu à plusieurs communications dans des séminaires et à l'encadrement d'un TER.

La phylogénie moléculaire a pour objectif de reconstituer l'histoire des êtres vivants au travers de celle de leur génome (ADN, ARN par exemple). L'évolution d'une séquence se fait par différents mécanismes de mutations aléatoires (insertion, délétion, substitution, duplication et transposition) qui changent la structure des gènes au cours du temps, et derrière lesquelles intervient un processus de sélection. Les séquences évoluent donc dans le temps le long des arêtes d'un arbre (dit phylogénétique), à partir d'une séquence ancestrale commune jusqu'aux séquences actuelles qui sont seules observables (voir un exemple sur la figure 5.1).

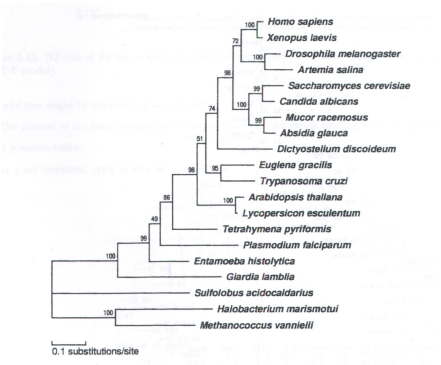


FIGURE 5.1 – Un exemple d'arbre phylogénétique, dont les feuilles sont les taxons

Les objectifs sont multiples, en classification du vivant pour comprendre le big bang de la vie, en génomique comparative pour interpréter des corrélations avec les caractéristiques morphologiques, ou en génomique fonctionnelle pour relier la variabilité évolutive des

positions à des changements de fonction, plus généralement pour comprendre l'évolution des génomes.

Une séquence est un ensemble de sites ordonnés du génome, à valeurs dans un alphabet fini $\mathcal{E}$: $\mathcal{E}$ est l'ensemble $\{A, T, G, C\}$ pour l'ADN, $\{A, U, G, C\}$ pour l'ARN et l'alphabet des vingt acides aminés pour les protéines. Les séquences de S taxons (feuilles de l'arbre phylogénétique) sont alignées par un processus bio-informatique permettant de faire correspondre les sites. Un alignement est le résultat de cette opération, il contient S séquences de n sites.

À partir de l'observation d'un alignement, deux problèmes d'estimation se posent : d'une part, la reconstruction de la topologie en cherchant un arbre phylogénétique T parmi l'ensemble des arbres à S feuilles ; d'autre part, l'estimation des paramètres associés à cette topologie : la longueur de branche et les paramètres propres au modèle choisi (taux relatifs de substitution des différents nucléotides par exemple).

Modélisation Les méthodes de reconstruction peuvent être déterministes (critère de parcimonie ou méthodes de distance) ou basée sur un modèle d'évolution (méthodes de maximum de vraisemblance ou méthodes bayésiennes), voir quelques exemples figure 5.2.

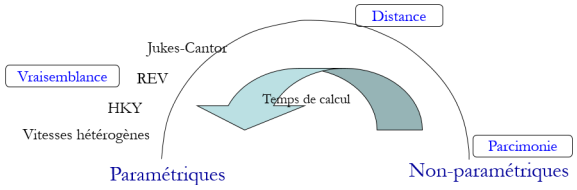


FIGURE 5.2 – Quelques méthodes de reconstruction en phylogénie

La modélisation probabiliste de l'évolution des séquences définit la loi $\mathbb{P}_{x\to y}(t)$ des substitutions dans le temps. Plus précisément, l'évolution d'un caractère $\{X_t, t \geq 0\}$ en un site le long d'une branche est un processus de Markov à temps continu, homogène

$$\forall t < u, \mathbb{P}(X_u = y | X_s = x, s < t) = \mathbb{P}(X_u = y | X_t = x)$$

Ce processus est caractérisé par une matrice de taux de substitution Q et une loi initiale π_0. La matrice de transition pour un temps fini t est telle que

$$\forall t \geq 0, \mathbb{P}(X_t | X_0) = \exp(Qt).$$

La longueur de branche définit le nombre moyen de substitutions sur l'arête correspondante. Différents modèles ont vu le jour : en supposant les sites sont indépendants, Jukes et Cantor (1969) proposent $Q_{ij} = \alpha$ pour $i \neq j$, et Yang, Adashi et Hasegawa (1985) une matrice de substitution symétrique. Yang (1994) définit le modèle *Rate Across Sites* qui permet un taux de substitution différent en chaque site h : $Q_h = r_h Q$ où r_h est tiré suivant une loi Gamma. L'estimation se fait par maximum de vraisemblance. Mais ces modèles supposent une vitesse d'évolution constante dans le temps, qui est contredite par les données réelles (Fitch (1970), Lopez, Casane et Philippe (2001)) : plus de 95% des positions variables du cytochrome b sont hétérotaches. Quantifier l'hétérotachie permettrait d'améliorer la détection d'homologues. Le modèle *covarion*, proposé par Fitch et Markowitz (1970) et formalisé par Tuffley et Steel (1998), permet de définir deux régimes de vitesse ON/OFF : lorsque le site est en régime ON, le processus est un processus de Markov de substitution Q, sinon, le site est invariable

— Le régime caché $\{S_t\}$ est une chaîne de Markov à temps continu de matrice de substitution R

$$R = \begin{pmatrix} -s_1 & s_1 \\ s_2 & -s_2 \end{pmatrix}; \ \mathbb{P}(S_t = v|S_0 = u) = [\exp(Rt)]_{u,v} := A(u, v|t)$$

— conditionnellement à $\{S_t\}$, X_t est généré suivant une chaîne de Markov à temps continu de loi initiale $b(u, i) = \mathbb{P}(X_0 = i|S_0 = u)$ et de matrice de transition

$$B(u, v, i, j|t) = \mathbb{P}(X_t = j|X_0 = i, S_t = v, S_0 = u)$$

Ainsi, la matrice de transition Π de la chaîne étendue (X_t, S_t) est

$$\Pi_t((u, i), (v, j)) = A(u, v|t)B(u, v, i, j|t)$$

$$\text{avec } \Pi_t = \exp(Mt) \text{ et } M = \begin{pmatrix} Q - s_1 & s_1\mathbb{I} \\ s_2\mathbb{I} & -s_2\mathbb{I} \end{pmatrix}.$$

Les variables cachées sont les régimes ON/OFF et les caractères aux noeuds internes de l'arbre. Les paramètres sont ceux des matrices R et Q, les lois initiales et les longueurs de branches.

Estimation D'une part, il est nécessaire de normaliser Q pour éviter une non-identifiabilité évidente. De plus, même dans ce cas, on ne peut espérer identifier qu'un arbre non raciné pour un modèle standard (homogène, réversible stationnaire). Mais il est possible de déterminer ses longueurs de branche même si elles n'ont pas une expression simple :

$$\begin{aligned} \mathbb{P}_{n_1,n_2}(i, j) &:= \mathbb{P}(X_{n_2} = j|X_{n_1} = i) \\ &= \mathbb{E}[\mathbb{P}(X_{n_2} = j|X_{n_1} = i; T_{n_1,n_2})] = \mathbb{E}[\exp(T_{n_1,n_2}Q)_{i,j}]. \end{aligned}$$

Le calcul de la vraisemblance se fait récursivement (Felsenstein, 1981) et j'ai développé en C++ le programme PMCov [1], qui, à partir des observations d'un alignement de séquences, estime par maximum de vraisemblance pour une topologie connue les différents paramètres du modèle et les temps de divergences entre les espèces.

Test d'une bi-partition La topologie d'un arbre décrit les événements de spéciation de S séquences. Elle peut être spécifiée par l'ensemble des bi-partitions induites par les branches de l'arbre. Avec Marie-Anne Poursat, nous avons étudié par simulation le niveau d'un test du rapport de vraisemblance d'une bi-partition induite par une branche identifiée a priori contre les bipartitions alternatives, dans le cas d'un alignement de $S = 5$ taxons. La statistique de test proposée est

$$RV = 2[\max_{j \in H_1} L_j - \max_{i \in H_0} L_i]. \tag{5.1}$$

Dans ce cas, H_0 et H_1 sont composites et leur composition est représentée figure 5.3.

Des simulations ont été conduites pour des alignements de taille $n = 1000$ avec le modèle HKY, une variante d'un modèle homotache [2]. Les séquences ont été générées par Seq-Gen [3] suivant les spécifications indiquées sur la figure 5.3. Elles n'ont pas permis de

1. `https://www.math.u-psud.fr/~keribin/PMCovCKN.htm`
2. `https://www.math.u-psud.fr/~keribin/PMCovCKN.htm`
3. `http://tree.bio.ed.ac.uk/software/seqgen/`

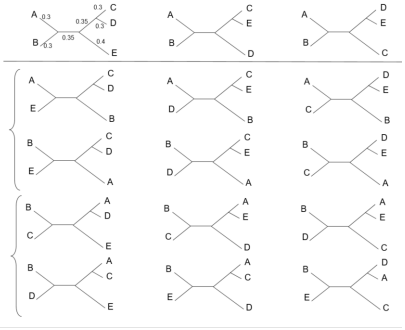


FIGURE 5.3 – Configurations topologiques sous H_0 (au dessus du trait), et sous H_1 (sous le trait), dans le cas du test d'existence de la bi-partition $(ab|cde)$. Valeurs des longueurs de branches pour le modèle de génératif sous H_0

faire apparaître de loi connue pour la loi de la statistique de test. Le seuil du test a été calculé par bootstrap avec $B = 500$ réplications, taille choisie après une étude de convergence. Le test s'avère très puissant, voire de puissance 1 quand les longueurs de branche sont égales. Une estimation de son niveau a montré sa cohérence, même si ce seuil est très variable en fonction de l'alignement. On aurait pu refaire l'étude avec des longueurs de séquences plus grandes, par exemple, $n = 3000$.

Par ailleurs, l'estimateur du maximum de vraisemblance sous H_1 d'un alignement généré sous H_0 peut renvoyer une topologie dont la branche $(ab|cde)$ est une longueur nulle. Il faudrait donc inclure les arbres non résolus dans les deux hypothèses, ce qui pose le problème d'une reformulation de la statistique de test. Ainsi, ce travail a pu montrer l'utilité du bootstrap pour le test de bi-partition en phylogénie, et permettait de nombreuses perspectives de recherche, mais il m'a pas pu être poursuivi faute de temps.

5.2 Feux de forêts

Ce travail a été réalisé en collaboration avec Charles Hernandez lors de sa thèse au Laboratoire de Météorologie Dynamique de l'école Polytechnique, encadrée par Philippe Drobinski, et a donné lieu à une publication (Hernandez et al., 2015).

Nous avons évalué les probabilités conditionnelles d'incendies (surface brûlée ou pouvoir radiatif) par rapport à des covariables météorologiques choisies (anomalie de température de 2 mètres, vitesse du vent de 10 mètres et anomalie de survenue des précipitations de janvier à juin).

Contexte Le *Canadian Fire Index* (Van Wagner, 1987) ou le système européen EFFIS permettent de définir des indices calibrés empiriquement pour expliquer si des conditions atmosphériques et hydrologiques sont sujettes au développement de feux. Cependant, l'un de leur principal inconvénient est leur manque de contraste temporel : ils identifient correctement les saisons propices au feu mais ne fournissent pas de variabilité à court terme du risque d'incendie (San-Miguel-Ayanz et al. 2013). Des modèles physiques ont aussi été implémentés, mais sont très exigeants en calcul et nécessitent une connaissance précise des conditions initiales et limites. (Preisler et al., 2004) ont utilisé un cadre probabiliste pour reconstituer les probabilités d'apparition et de propagation d'incendies importants à l'aide de covariables météorologiques et géographiques. Les résultats, bien qu'encourageants, n'ont donné qu'une qualité mitigée de l'estimation de la fréquence mensuelle des

incendies. Nous avons décidé d'étudier par des méthodes statistiques la *surface brûlée* (BA) et le *pouvoir radiatif* du feu (FRP) qui sont deux variables importantes déterminant l'intensité d'un feu

Données Les données BA et FRP sont issues d'images de spectrométrie de moyenne résolution (MODIS) et concernent des observations du bassin méditerranéen. Deux jeux BAM et BAE ont été dérivés pour BA, suivant le type de résolution utilisée : (1) le jeu de données BAM est bâti sur une résolution de 10 km choisie comme un bon compromis pour conserver des informations suffisamment détaillées sur le lieu de l'incendie et faciliter la comparaison avec les données météorologiques provenant de ERA-Interim ; (2) le jeu de données BAE, enregistrant des surfaces de 40 ha avec une résolution plus fine. Un système de localisation 3D permet l'agrégation éventuelle de grandes parcelles incendiées qui sont susceptibles d'être le plus influencées par les conditions météorologiques. Ces données, déjà prétraitées par Charles Hernandez, comportent 5821 observations de feu pour BAM, 4840 pour BAE et 24 273 pour FRP. Comme leur répartition est très asymétrique, nous les avons systématiquement utilisées après l'application d'une fonction logarithme.

Estimation Une première approche utilisant des techniques de régression (classique ou par réseaux de neurones) en fonction d'une vingtaine de variables météorologique s'est soldée par un échec, et nous avons choisi de recentrer notre analyse sur l'estimation d'une loi conditionnelle calculée en certains points d'une grille de valeurs de trois variables météorologiques choisies : anomalie de température à 2 mètres au dessus du sol, vitesse du vent à 10 mètres au dessus du sol et anomalie dans la survenue des précipitations de janvier à juin par rapport à une année moyenne. Ce choix de variables permet de conserver un large éventail d'échelles de temps, d'utiliser des variables ayant un impact prouvé sur l'activité des incendies de forêt, et de limiter le problème liée à la dimension.

Parmi un ensemble de lois paramétriques (dont les lois de Gamma, Cauchy, exponentielle tronquée), c'est la loi log-normale qui offre le meilleur l'ajustement pour chacune des variables $\log_{10}(BAE)$, $\log_{10}(BAM)$ et $\log_{10}(FRP)$. Soit $f_{Y|\mathbf{x}}(y)$ la densité d'une variable Y d'intensité de feu conditionnellement à trois variables météorologiques notées $\mathbf{x}$. Y peut représenter $\log_{10}(BAE)$, $\log_{10}(BAM)$ ou $\log_{10}(FRP)$. En chaque point $\mathbf{x}$ d'une grille de conditions d'expérience, nous avons estimé le paramètre de la loi log-normale conditionnelle en retenant 10% des points de l'échantillon les plus proches de $\mathbf{x}$, ce qui est suffisant pour l'estimation. Les intervalles de confiance ont été générés par bootstrap.

La figure 5.4 représente les contours de la probabilité estimée d'un incendie particulièrement grand (car dépassant le seuil de 1000 ha), calculé à partir de notre méthode. On observe que la probabilité d'un incendie de grande surface brûlée est une fonction croissante de la vitesse du vent. Celle-ci augmente avec la température à 2 m ou le vent à 10 m particulièrement quand il y a un déficit de précipitation. On retrouve deux modes de surface brûlée suivant la vitesse du vent qui ont été discutés et analysés par Hernandez (2015). Des résultats similaires sont obtenus avec BAM et FRP.

Le modèle appliqué aux conditions météorologiques du plus grand incendie arrivé à l'été 2003 en Espagne (BA = 260 000 ha, FRP = 731 MW) permet de suivre dans le temps de l'été la probabilité de survenue d'un incendie de plus de 1000 ha ou d'intensité supérieure à 1500 MW (qui correspond au seuil d'intensité où le nuage de fumée atteint la troposphère libre). La figure 5.5 présente les résultats. Deux lignes noires indiquent le début et la fin de l'incendie. Pendant l'incendie, la probabilité d'incendies impliquant de une grande aire brûlée est significativement supérieure et atteint 13%, alors qu'elle reste à

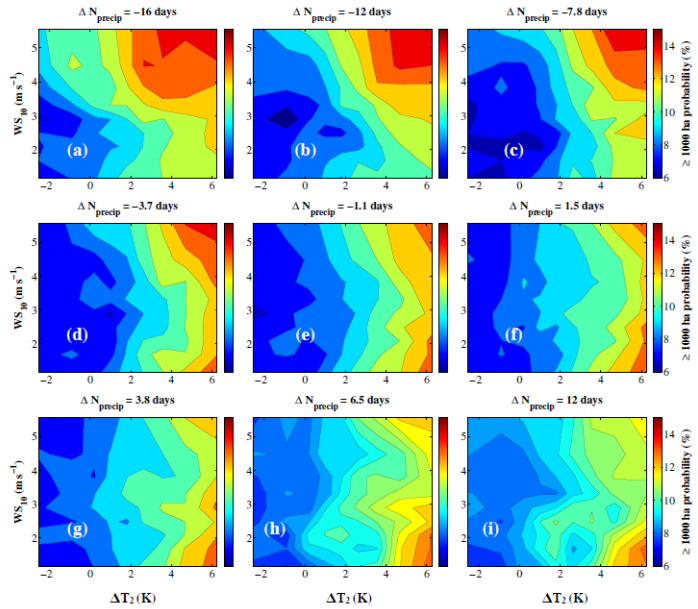


FIGURE 5.4 – Probabilité estimée d'observer une surface brûlée de plus de 1000 *ha* (données BAM), en fonction de la température (en abscisse) et la vitesse du vent (en ordonnée) pour 12 régimes de précipitation.

environ 8% le reste du temps. La probabilité d'incendies de forte puissance triple, même si elle reste faible, et la variation est significative. Les incertitudes sur les caractéristiques météorologiques ont été étudiées et sont assez faibles, autour de 2% pour BA et 0.05% pour RFP.

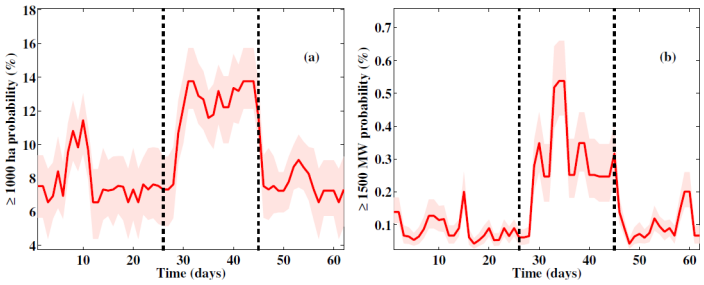


FIGURE 5.5 – A partir du jeu de données BAM, probabilités d'observer un incendie de plus de 1000 *ha* (à gauche) et d'intensité supérieure à 1500 *MW* (à droite) en fonction du temps pour la période de juillet à août 2003, calculées avec les covariables du plus grand feu de forêt qui ait eu lieu au Portugal cette saison-là. Les lignes pointillées noires indiquent le début et la fin de l'événement. En rouge légèrement ombré, les intervalles de confiance à 90%.

En conclusion, notre méthode est prometteuse pour identifier des périodes où le risque d'incendie est relativement plus élevé, et une telle approche probabiliste pourrait être la base d'un futur système d'alerte au risque incendie.

5.3 Agrégation d'experts en finance

J'ai été contactée en 2015 par Jean-Marc Guillard (Dassault Systèmes) dont j'avais été la collaboratrice. Ayant développé à titre personnel un algorithme d'aide à l'investissement boursier (ST4), il souhaitait réfléchir à la possibilité d'utiliser l'apprentissage automatique pour l'améliorer et discuter de l'opportunité de la création d'une start-up [4]. Ces travaux ont mené à la réalisation d'un stage de master que j'ai encadré et à une communication dans un congrès à l'étranger en 2017. Nous avons comparé les stratégies standards à la stratégie ST4, puis avec celles d'agrégation d'experts, avec des expérimentations basées sur les prix des actifs du CAC40 de 2014 à 2017.

Contexte La modélisation des marchés financiers est un sujet qui peut s'étudier sous plusieurs angles : modélisation probabiliste (modèle de *Black and Scholes* où le prix d'une action est un processus stochastique à temps continu, utilisé pour les produits dérivés et la couverture de risque), l'approche économétrique (séries temporelles de plus en plus sophistiquées : modèles GARCH, ARFIMA par exemple). Mais ces modèles peuvent se révéler irréalistes soit à cause de forte condition sur la structuration du marché ou d'hypothèses mathématiques trop réductrices. Des approches en apprentissage statistique ont proposé l'utilisation de SVM ou de réseaux de neurones. La variable d'étude est le cours de l'action P_t, son rendement (*return*)

$$S_t = \log(P_t/P_{t-1})$$

ou une variable binaire de tendance (à la hausse, à la baisse $T_t = \mathbb{1}_{S_t>0}$). Arroyo (2011) a étudié la prévision de séries temporelles d'intervalles (court le plus haut et le plus bas dans la journée) en comparant des techniques univariées et multivariées permettant la cointégration et a montré que ces dernières étaient adaptées pour la prédiction de l'étendue de la série, ouvrant une autre voie à la prévision de la volatilité.

Stratégies La prédiction du cours d'un actif est bien évidemment intéressante pour l'investisseur, mais il lui faut une stratégie (qu'acheter, que vendre, à quel moment) pour l'aider à optimiser son portefeuille. Un portefeuille avec m actifs est défini au temps t, par le vecteur de portefeuille b_t :

$$b_t = (b_{t,1}, ..., b_{t,m}) \in \{b \quad | \quad b_{t,i} \geq 0, \sum_{i=1}^{m} b_{t,i} = 1\}$$

où $b_{t,i}$ est la proportion de capital investi dans l'actif i. La *stratégie* est une suite de b_t, définie sur la fenêtre de marché $t = 1, \ldots, n$. La *richesse cumulée* est

$$S_n(b) := S_0 \prod_{t=1}^{n} < x_t, b_t >= S_0 \times CR_n(b)$$

où x_t est le vecteur des *prix relatifs* entre les prix de clôture en t et $t-1$, et S_0 est l'investissement initial. La sélection séquentielle de portefeuille (*online portfolio selection*) sélectionne un portefeuille de façon séquentielle parmi un ensemble d'actifs : la décision d'allocation est prise immédiatement à chaque période de négociation t : l'objectif est de maximiser le *rendement cumulé* à l'issue du nombre fixé n de périodes :

$$\max_b CR_n(b) = \max_b \log CR_n(b).$$

4. `http://www.stats4trade.com/`

Il existe un grand nombre d'algorithmes de sélection (Li et Hoi, 2014) : approches d'achat et de maintien (*Buy and Hold* (BAH)), de réallocation systématique (*Constant Rebalanced Portfolios* (CRP) : $b_t = b$), de meilleure performance passée (optimale BAH : $b^* = \arg\max_b < b, \otimes_{t=1}^n x_t >$ où $(x \otimes y)_i = x_i y_i$; optimal CRP : $b^* = \arg\max_b \sum_{t=1}^n \log(< b, x_t >)$). La stratégie de suivi du gagnant *Follow the Winner* augmente la proportion d'actifs les plus performants, comme c'est le cas pour la méthode du gradient exponentiel :

$$b_{t+1,i} \propto b_{t,i} \exp\left(-\eta \quad \frac{-x_{t,i}}{< b_t, x_t >}\right), \quad \eta > 0.$$

Le pas η est choisi a priori et constant, ou calculé sur une fenêtre d'étalonnage de longueur ω à déterminer

$$\eta_t = \arg\max_\eta \prod_{\tau=\omega}^{t} < x_\tau, b_\tau(\eta) >, \quad 1 \leq \omega \leq t.$$

Citons également la stratégie *Follow the Loser* qui augmente la proportion des actifs les moins performants (Passive Aggressive Mean Reversion (PAMR) par exemple) ou l'approche *Pattern Matching* construisant sa stratégie sur un cluster de prix historiques similaires (Nonparametric Kernel-based Log-optimal Strategy (NKLS)). La performance de ces algorithmes peut dépendre de la tendance du marché, certains étant plus adaptés en période de baisse ou de hausse. La stratégie S4T de Jean-Marc Guillard est basée sur des techniques d'apprentissage supervisé. Elle procède en trois étapes :

1. la prédiction à la hausse ou à la baisse à 20 jours en utilisant un algorithme de bagging, combinant les résultats d'arbre de classification construits par bootstrap sur un ensemble d'apprentissage. Une cinquantaine de variables explicatives caractérisant les cours ont été définies mais restent confidentielles, 500 arbres bootstrap sont générés. La décision est faite à la majorité, et un indice de confiance généré.
2. investissement séquentiel d'un ratio du montant total en fonction d'un seuil de l'indice de confiance à 20 jours
3. Une ligne est conservée pendant 20 jours avant d'être remise en jeu.

Le processus est piloté par le montant à investir, le nombre de lignes d'investissement, et un ratio maximum pour les composantes du vecteur de portefeuille. La figure 5.6 montre une comparaison entre les différentes stratégies : la longueur de la fenêtre d'étalonnage comme la performance peuvent dépendre de la tendance du marché et on observe que S4T est un bon compromis, aussi bien à la hausse qu'à la baisse, et moins bruité.

Agrégation d'experts Suivant Stoltz et Lugosi (2005), nous avons mis en place une stratégie d'agrégation d'experts pour combinant à chaque pas D stratégies (experts). Soit P_t le vecteur de poids au temps t. Pour $d = 1, \ldots, D$,

$$P_t(d) \geq 0, \quad \sum_{d=1}^{D} P_t(d) = 1$$

Plusieurs méthodes peuvent être utilisées. L'algorithme d'agrégation (AA) met à jour le poids d'un expert d suivant

$$P_{t+1}(d) \propto P_t(d) \exp(-\eta \quad \ell(x_t, b_t(d))), \quad \eta > 0$$

avec $\ell(x_t, b_t(d)) = -\log(< x_t, b_t(d) >)$, soit

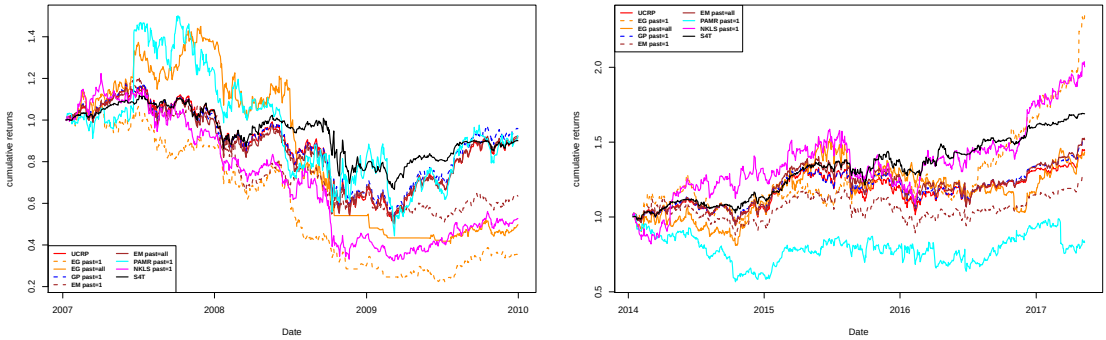


FIGURE 5.6 – Rendement de différentes stratégies en période de baisse (à gauche) et en période de hausse (à droite). ST4 est un trait noir plein. PAST indique la longueur de la fenêtre d'étalonnage (all pour tout le passé, 1 pour la dernière période

$$P_{t+1}(d) \propto P_t(d) < x_t, b_t(d) >^{\eta}$$

L'algorithme d'agrégation simplifié (SAA) pose $P_{t+1}(d) \propto< x_t, b_t(d) >^{\eta}$.

C'est ce dernier qui a été retenu. L'effet stabilisateur de l'agrégation (plus d'experts, moins de risques) est bien observé sur la figure 5.7, avec une interrogation sur son comportement à la baisse laisse à penser un problème de calibration interne. Nous avons retenu de cette étude que le machine learning peut aider à la définition de stratégies séquentielles de portefeuille, et que la solution S4T se révélait compétitive, ce qui répondait à une interrogation initiale. Des directions de recherche ont été définies, comme l'affinage de la définition de la fenêtre d'étalonnage, ou son affranchissement en utilisant des algorithmes d'agrégation ne nécessitant pas de paramètres d'étalonnage (ML prod, ML poly 3, (Gaillard et Goude, 2015)) ; enfin, le pilotage du montant à investir, du nombre de lignes d'investissement et du ratio maximum qui sont actuellement défini a priori par l'investisseur pourraient tirer avantageusement partie de méthodes d'apprentissage automatique. Mais nos contraintes respectives n'ont pas pu nous permettre de dégager du temps pour la poursuite de cette étude.

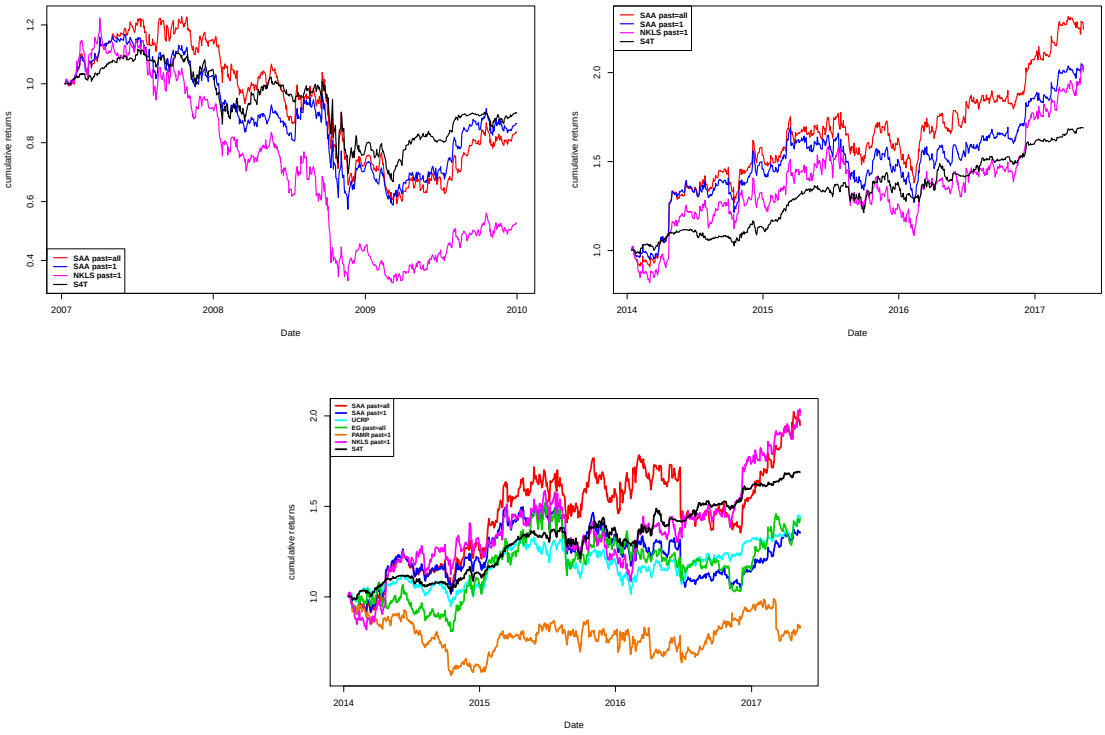


FIGURE 5.7 – Agrégation des deux stratégies NKLS et S4T sur la période 2007-2010 (à gauche) et 2014-2017 (à droite), et de cinq stratégies sur la période 2014-2017 (en bas)

Chapitre 6

Projet de recherche

Le co-clustering en général, et les modèles LBM et SBM en particulier, sont des domaines toujours particulièrement féconds. Il suffit d'en regarder le nombre de publications récentes, que ce soit pour des lois conditionnelles de plus en plus complexes (loi binomiale négative (Aubert, 2017), données ordinales (Jacques et Biernacki, 2018), données fonctionnelles (Slimen et al., 2018)) ou pour un MLBM avec des lois de types différents (Selosse et al., 2018). On pourrait envisager de prendre en compte des co-variables dans la loi conditionnelle. Bergé et al. (2019) proposent de combiner le LBM et l'allocation de Dirichlet latente (LDA), afin de regrouper simultanément des lignes et des colonnes de données d'interaction textuelle. Le SBM se décline maintenant en SBM-graphon (Latouche et Robin, 2013), en STBM (SBM dans un contexte dynamique où un réseau est observé au fil du temps avec des créations et des destructions d'arêtes, Rastelli (2019)) ou en blocs chevauchants (Latouche et al., 2011). Caron et Fox (2017); Caron et Rousseau (2017) ont proposé un nouveau modèle de graphe sparse, et Massoulié (2014) a étudié des graphes de transition de phase afin de déterminer les limites de détectabilité d'un SBM.

Mes travaux précédents ont illustré la grande parcimonie des méthodes de clustering, qu'elles soient pour partitionner les variables dans un schéma supervisé ou dans le schéma doublement non supervisé de la classification croisée. Par ailleurs, des points restent ouverts et des interrogations demeurent. Mon projet de recherche est bâti sur ces fondations, et abordera des points méthodologiques, théoriques et appliqués.

Les données massives fournissent un terrain toujours particulièrement intéressant pour le co-clustering. Trois directions de recherche méthodologique peuvent en découler : (1) montrer que le co-clustering peut être une méthode parcimonieuse efficace pour effectuer du clustering en grande dimension (6.1.1) ; (2) trouver des solutions au passage à l'échelle des données de pharmacovigilance (6.1.2) ; (3) plus largement, proposer une méthodologie adaptative de quantification des données en grande dimension (6.1.3), permettant d'aborder le compromis entre précision statistique et ressources informatiques.

Le LBM propose encore des défis théoriques, par exemple : (1) montrer la consistance des critères BIC et ICL (6.2.1) permettant de répondre à la conjecture de Keribin et al. (2015) ; (2) dériver une heuristique de nappe pour le LBM (6.2.2) ; (3) plus largement, étudier les critères non asymptotiques de choix de modèle quand tous les modèles sont faux (6.2.3), répondant ainsi à une interrogation de Keribin (2019).

Enfin, je suis toujours à l'écoute d'opportunités applicatives, comme par exemple récemment avec la mise en place de deux projets de thèse : l'un avec la SNCF sur la modélisation et la prévision des temps d'échanges en gare (6.3.1), l'autre avec PSA la construction d'un critère probabilisé de fatigue multiaxiale tenant compte du comportement aléatoire de la structure et du matériau (6.3.2).

6.1 Co-clustering et données de grande dimension

6.1.1 La classification croisée comme outil de classification en grande dimension [LBM, grande dimension]

En collaboration avec *Christophe Biernacki (INRIA-Modal)*.

Les modèles standards de classification non supervisée sont connus pour être très efficaces sur des données de faible dimension, c'est-à-dire quand il y a peu de variables, mais ils souffrent de limitations à la fois statistiques et informatiques en grande dimension.

Afin de contrebalancer ce fléau de la dimension, certaines propositions ont été faites pour prendre en compte la définition de motifs et la redondance, mais les modèles correspondants ne supportent pas un trop grand nombre de variables : voir par exemple l'algorithme SelvarClust de Maugis et al. (2009) pour la sélection de variables en classification d'observations gaussiennes ou récemment Fop et Murphy (2018) pour l'analyse de classes latentes. Laclau et Brault (2019) ont proposé de définir une classe de variables non pertinentes dans le LBM.

Nous pensons que la classification croisée présente un intérêt particulier pour la classification d'observations en grande dimension, même si ce n'est pas sa mission primaire, comme illustré dans les applications de neuro-imagerie (section 4.4). En effet, le regroupement des colonnes est ici reformulé en tant que stratégie de contrôle de la variance de l'estimation, la dimension du modèle étant déterminée par le nombre de groupes de variables au lieu du nombre de variables lui-même. Cependant, la contrepartie statistique de cette réduction importante de la variance apporte naturellement un biais important.

Nous souhaitons étudier, aussi bien d'un point de vue théorique que pratique, le comportement biais-variance de la stratégie de classification croisée pour la classification non supervisée, dans des scénarii présentant les caractéristiques d'un jeu de données de grande dimension (variables corrélées, variables non pertinentes par exemple). Ceci passe par la nécessité de préciser la notion de redondance et de pertinence d'une variable ou d'un ensemble de variables vis-à-vis du processus de classification. Nous souhaitons montrer la capacité de la classification croisée à surpasser la simple classification non supervisée des lignes, même si la classification croisée correspond clairement à une situation de modèle mal spécifié, révélant ainsi une manière prometteuse de traiter efficacement la classification en (très) grande dimension.

6.1.2 Méthodologie de grands jeux de données [MLBM, grande dimension, sparsité]

En collaboration avec *Gilles Celeux (INRIA-*Celeste*)*.

La thèse de V. Robert s'est achevée sur le challenge posé par la taille des données individuelles ($200000 \times (1500 + 2000)$), qui pose un vrai problème méthodologique. Les algorithmes actuellement proposés pour estimer un LBM nécessitent d'être adaptés pour passer à l'échelle de très grands jeux de données. Il s'agit de comparer plusieurs techniques comme l'utilisation d'un cluster de calcul, l'adaptation de l'algorithme en utilisant les matrices à stockage parcimonieux, la mise en œuvre de techniques Map/Reduce (partitionnement du jeu de données, détermination d'une partition croisée sur chaque sous-ensemble, puis réconciliation des partitions obtenues) qui pourront mener à repenser complètement le principe des algorithmes d'estimation. Dans un second temps, l'étude expérimentale, puis théorique, de techniques comme le sous-échantillonnage ou le regroupement dans un contexte de données clairsemées permettront d'étudier la relation entre la précision statistique et la complexité algorithmique.

6.1.3 Grands jeux de données et ressources limitées [grande dimension, compromis précision-ressources]

En collaboration avec *Christophe Biernacki (INRIA-Modal)*, co-direction d'une thèse débutant en octobre 2019.

De façon plus générale, l'afflux constant de données, souvent créées par des processus automatiques comme des capteurs spécialisés dans le monde réel ou l'acquisition de données Web dans le monde virtuel, rend nécessaire le réglage automatique d'un compromis entre la précision statistique qu'il est possible d'atteindre et les ressources informatiques disponibles dans le respect des limites physiques.

Le projet que nous souhaitons mener consiste à adapter le paradigme général de l'approche modèle en clustering au cas de grands ensembles de données, en accordant une attention particulière aux situations de grande dimension. La solution que nous souhaitons développer consiste à *quantifier* l'ensemble de données, c'est-à-dire réduire un jeu de données initial de N individus en un nouvel ensemble de données de $n < N$ individus pondérés. Dans le cas continu multidimensionnel, cela correspond généralement à la construction d'intervalles multidimensionnels (sous forme d'une grille) conduisant à ce que l'on appelle des données regroupées ou *binned data* (Samé et al., 2006; Wu, 2014).

Nous souhaitons comparer cette méthode avec des solutions plus classiques comme le sous-échantillonnage, en particulier sur le terrain de la capacité théorique à détecter des groupes de faible fréquence. Cette propriété est particulièrement importante dans la mesure où l'intérêt de collecter d'énormes jeux de données est souvent guidé par la possibilité de détecter un nouveau signal, donc peu fréquent mais potentiellement important. En réduisant considérablement la taille du jeu de données de N à n (avec pondérations), la mémoire, l'énergie et le temps de calcul requis peuvent être considérablement réduits. Cependant, la grille définie dans un espace de grande dimension ne peut plus être régulière à cause de la malédiction de la dimension (paradigme du *vide*, Bouveyron et Brunet-Saumard (2014)). Par conséquent, la question principale à traiter dans ce projet est l'automatisation de la construction d'une grille adaptative permettant de préserver la capacité à détecter de "petits" clusters d'une part et la consistance de l'estimation sur une grille irrégulière d'autre part. L'estimation de cette grille adaptative permettra d'étudier le compromis entre précision statistique et ressources informatiques disponibles, étendant les résultats de grille régulière déjà obtenus en petite ou moyenne dimension (Samé et al., 2006; Wu, 2014).

6.2 Sélection de modèle [LBM, Théorie]

6.2.1 Consistance des critères asymptotiques de choix de modèle [LBM, critères asymptotiques]

L'étude théorique de l'asymptotique du LBM a été conduite lorsque le nombre de blocs $(K \times L)$ est connu. Son extension naturelle est l'étude du cas où le nombre de blocs est inconnu. C'est le chemin suivi par Wang et al. (2017) qui ont étudié l'asymptotique du rapport de vraisemblance d'un modèle SBM quand le nombre de blocs est mal spécifié. Je souhaiterais étendre ces travaux au cas de la double asymptotique (ligne, colonne) du LBM et démontrer la consistance d'un critère de vraisemblance pénalisée.

Par ailleurs, je souhaiterais également prouver la conjecture de Keribin et al. (2015) prétendant le comportement asymptotiquement équivalent des critères BIC et ICL pour le LBM. La consistance du critère ICL serait propre au LBM, elle n'est bien sûr pas vérifiée

dans le cas des mélanges simples. En effet, BIC et ICL sont liés asymptotiquement par la décomposition

$$ICL(\mathbf{x}, \mathbf{z}, \mathbf{w}; K, L) = BIC(\mathbf{x}; K, L) - \sum_{ijkl} p(z_{ik} w_{j\ell} | \mathbf{x}; \widehat{\theta}_{KL}) \log p(z_{ik} w_{j\ell} | \mathbf{x}; \widehat{\theta}_{KL})$$

où apparaît un terme d'entropie de la loi des labels conditionnellement aux observations. Le critère ICL est lui même calculé en remplaçant les labels inconnus $(\mathbf{z}, \mathbf{w})$ par leur estimation $(\widehat{\mathbf{z}}, \widehat{\mathbf{w}})$ résultant de la règle du maximum a posteriori (MAP). Cette estimation est consistante pour la vraie partition quand le nombre de blocs est connu, et l'entropie s'annule. Les deux critères calculés sous le bon nombre de blocs sont donc asymptotiquement équivalents. Il serait donc intéressant de déterminer si les critères ICL et BIC sont également asymptotiquement équivalents quand le nombre de blocs du modèle est mal spécifié, ou à défaut, si ICL est lui aussi consistant.

6.2.2 Critère de choix non asymptotique [LBM, non asymptotique]

Le critère ICL étant calculable de façon exacte à distance finie dans le modèle LBM, il est intéressant de le comparer avec un critère issu d'une inégalité oracle. Si l'inégalité définit la forme des pénalités, il faudra en calibrer les constantes pour pouvoir utiliser en pratique les critères pénalisés associés. Nous pourrons envisager la méthode de l'heuristique de pente [Birgé et Massart 2001, 2006] pour déterminer des pénalités optimales à partir des données. Mais de part la double direction (nombre de classe en lignes nombre de classe en colonne), il faudra l'étendre à une heuristique de nappe, à l'image de ce qui est fait dans la thèse de Michel [2008].

6.2.3 Choix de modèle quand tous les modèles sont faux [non asymptotique]

En collaboration avec *Jean-Patrick Baudry (UPMC).*

Dans l'étude de la définition d'une carte d'identité d'une tumeur cancéreuse (section 1.4) la méthode de sélection non asymptotique de l'heuristique de pente a surpassé le critère BIC. Le modèle considéré pour l'estimation était certainement biaisé, et nous avions formulé l'hypothèse que l'heuristique de pente pourrait être justifiée lorsque tous les modèle sont faux, mais que leur biais devient constant. Il s'agit ici de montrer cette conjecture.

Considérons une collection emboîtée de modèles $M_1 \subset \ldots \subset M_{m-1} \subset M_m$. La justification de l'heuristique de pente nécessite de supposer que le biais s'annule asymptotiquement pour un modèle de dimension assez grande, et pour un modèle de très grande dimension. C'est par exemple le cas en régression gaussienne homoscédastique à design fixe [Arlot 2011]. Nous souhaitons supprimer l'hypothèse de biais nul asymptotique pour considérer le cas où tous les modèles de la famille seraient faux, ce qu'il est raisonnable de considérer dans les cas pratiques.

6.3 Modélisation de données applicatives

6.3.1 Prévision de temps d'échange lors des stationnements de trains en gare [Machine learning supervisé]

En collaboration avec *Gilles Stoltz (LMO, Université Paris Sud)*, co-direction d'une thèse Cifre SNCF débutant en octobre 2019.

Un des facteurs de la ponctualité des trains en zone dense (et de la gestion de crise en cas d'incident sur une ligne) est le respect à la fois du temps de parcours entre deux gares et du temps de stationnement dans une gare. Un des minorants du temps de stationnement est le temps d'échange, déterminé par le temps mis pour faire descendre et monter tous les passagers qui le souhaitent. Il dépend, entre autres, du train, de sa mission, du calendrier, des horaires, de la charge instantanée et de la configuration du quai ou de la gare. Des études préliminaires internes à la SNCF ont montré que le problème était complexe (Cornet et al., 2018).

À partir d'un jeu de données concernant la ligne E du Transilien, nous adresserons la prédiction (apprentissage automatique) et la modélisation (statistique) : (1) construire une typologie propre gare-heure, gare-heure-type de train, par exemple avec des techniques de *co-clustering* ; (2) étudier les corrélations entre nombre de voyageurs (charge) et flux en gare, flux et temps de stationnement, et éventuellement d'autres variables à définir ; (3) modéliser les flux ou charges (au sein d'une même gare, ou d'un même train) comme un processus stochastique ; (4) développer un simulateur numérique réaliste des flux de voyageurs et tester différents scenarii d'incidents et de résolution, afin de proposer des intuitions de résolution efficaces.

6.3.2 Construction d'un critère probabilisé de fatigue multiaxiale [Apprentissage, méthodes bayésiennes]

Direction d'une thèse Cifre PSA débutant en novembre 2019, co-direction *Patrick Pamphile (LMO, Université Paris Sud).*

La digitalisation de la conception est au cœur des processus des départements métier des constructeurs automobiles, pour leur permettre de réduire les coûts et les temps de développement. Ceci s'applique également aux études de fiabilité de certains composants du châssis d'un véhicule, et la volonté est de réduire drastiquement le nombre d'essais physiques pour tendre vers une conception presque entièrement numérique n'ayant qu'une seule phase de validation.

Les modèles déterministes, bien que développés à partir de dessins de conception détaillée, peuvent prédire des comportements différents de ceux observés sur la structure lors d'essais. Ces écarts peuvent être dus à la discrétisation plus ou moins fidèle à la géométrie de la structure, aux incertitudes sur certains paramètres du modèle (tels que les propriétés des matériaux, les conditions aux limites), ou aux chargements aléatoires subis par la structure (Beck et Katafygiotis, 1998). Il est important de mettre à disposition de nouvelles méthodes en complément de la modélisation déterministe classique par éléments finis (EF), pour permettre d'exploiter le capital de données accumulées au cours des années sur l'ensemble des projets : résultats de calculs, mesures et données d'essais.

Un des objectifs de ce projet est de proposer une modélisation probabiliste du comportement d'une structure à partir d'une modélisation par EF, prenant en compte les fluctuations non assignables du modèle, afin de définir un critère probabiliste de rupture et ses marges de confiance. Les trois étapes suivantes sont envisagées : (1) Définir des lois a priori informatives pertinentes utilisant le retour d'expérience (REX) métier comme infor-

mation a priori et utiliser une estimation bayésienne pour calibrer les paramètres. Ce REX est conséquent et nécessitera des traitements statistiques avancés de *machine learning*, et en particulier en *clustering* pour identifier des similitudes ou des motifs proches parmi plusieurs modèles. L'estimation mettra en œuvre des méthodes bayésiennes non-itératives (Celeux et Pamphile, 2019), moins coûteuses et moins instables que les méthodes classiques, ce qui permettra de tester leur efficacité dans ce cadre. (2) Sélectionner les paramètres (physiques ou de modélisation) importants. (3) Définir un critère probabiliste de fatigue coaxiale tenant compte à la fois du comportement aléatoire de la structure et du matériau (Fouchereau et al., 2014) étendant les critères déterministes existants (Dang-Van, 1993).

6.4 Diriger des projets

En sus des connaissances acquises pendant mes études d'ingénieur, j'ai développé des compétences d'encadrement et de gestion de projet lors de mon expérience à Dassault-Systèmes, où j'ai dirigé des projets et équipes de différentes tailles, comptant jusqu'à une dizaine d'ingénieurs : aussi bien d'un point de vue technique (définition des projets et des objectifs, choix des solutions techniques, ordonnancement, ...), que humain (encadrement, suivi de carrière, résolution de conflits, ...) et de relations extérieures (projets avec des partenaires et clients français ou étrangers, lien avec la recherche, présentation des solutions dans des congrès d'utilisateurs). J'ai suivi dans ce cadre des formations au management adaptatif, à la gestion de projet, la gestion des conflits, celle des entretiens individuels et à la démarche qualité par exemple. J'ai également pu mettre en œuvre ces compétences pendant l'année (2005-2006) où j'ai dirigé le département informatique de l'IUT d'Orsay.

L'encadrement de l'enseignement par la recherche d'étudiants a commencé pendant ma période industrielle pendant laquelle j'ai encadré plusieurs stages à la frontière entre recherche et industrialisation. J'ai également encadré des stages d'introduction à la recherche du niveau L3 au niveau M2. Enfin, j'ai beaucoup appris sur l'encadrement de projet de recherche scientifique en co-dirigeant les thèses de Vincent Michel, Vincent Brault et Valérie Robert. Ainsi, il est temps pour moi de pouvoir pleinement prendre la responsabilité de mener des projets de recherche scientifique.

Bibliographie

E. Allman, C. Mattias, et J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37 :3099–3132, 2009.

C. Ambroise et C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(1) :3–35, 2012.

S. Arlot. Minimal penalties and the slope heuristics : a survey. *arXiv preprint arXiv :1901.07277*, 2019.

S. Arlot, A. Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4 :40–79, 2010.

J. Aubert. *Analyse statistique de données biologiques à haut débit*. PhD thesis, Université Paris-Saclay, 2017.

A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, et D. S. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8(Aug) :1919–1986, 2007.

J.-P. Baudry. *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes*. PhD thesis, Université Paris Sud-Paris XI, 2009.

M. J. Beal et al. *Variational algorithms for approximate Bayesian inference*. university of London London, 2003.

J. L. Beck et L. S. Katafygiotis. Updating models and their uncertainties. i : Bayesian statistical framework. *Journal of Engineering Mechanics*, 124(4) :455–461, 1998.

L. R. Bergé, C. Bouveyron, M. Corneli, et P. Latouche. The latent topic block model for the co-clustering of textual interaction data. *Computational Statistics & Data Analysis*, 2019.

P. Bickel, D. Choi, X. Chang, H. Zhang, et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4) :1922–1943, 2013.

C. Biernacki. *Model choice and model aggregation*, chapter Mixture Models. Editions Technip, 2017.

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

C. Biernacki, G. Celeux, et G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4) :561–575, 2003.

L. Birgé et P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268, 2001.

L. Birgé et P. Massart. Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138(1-2) :33–73, 2007.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

C. Bouveyron et C. Brunet-Saumard. Model-based clustering of high-dimensional data : A review. *Computational Statistics & Data Analysis*, 71 :52–78, 2014.

V. Brault. *Estimation et sélection de modèle pour le modèle des blocs latents*. PhD thesis, Université Paris Sud, 2014.

V. Brault et A. Lomet. Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau. *Journal de la Société Française de Statistique*, 156(3) :27–51, 2015.

V. Brault, G. Celeux, et C. Keribin. Régularisation bayésienne du modèle des blocs latents. Dans *44ème Journées de Statistique*, 2012.

V. Brault, G. Celeux, et C. Keribin. Mise en œuvre de l'échantillonneur de Gibbs pour le modèle des blocs latents. Dans *46èmes journées de statistique de la SFdS*, 2014.

V. Brault, C. Keribin, et M. Mariadassou. Consistency and asymptotic normality of latent block model estimators. *arXiv preprint arXiv :1704.06629v3*, 2019. URL `https://arxiv.org/pdf/1704.06629.pdf`. soumis.

F. Caron et E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(5) :1295–1366, 2017.

F. Caron et J. Rousseau. On sparsity and power-law properties of graphs based on exchangeable point processes. *arXiv preprint arXiv :1708.03120*, 2017.

M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, et A. R. Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1) :112–122, 2009.

G. Celeux et J. Diebolt. L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de statistique appliquée*, 34(2) : 35–52, 1986.

G. Celeux et G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3) :315–332, 1992.

G. Celeux et P. Pamphile. Competing risk with masked causes and highly censored data : Non-iterative estimations. Technical report, INRIA, 2019.

G. Celeux, D. Chauveau, et J. Diebolt. Stochastic versions of the EM algorithm : an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4) :287–314, 1996.

A. Celisse, J.-J. Daudin, et L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6 :1847–1899, 2012.

S. Chib et I. Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453) :270–281, 2001.

A. Clément, A. Riviere, P. Serré, et C. Valade. The TTRSs : 13 constraints for dimensioning and tolerancing. Dans *Geometric design tolerancing : theories, standards and applications*, pages 122–131. Springer, 1998.

S. Cornet, C. Buisson, F. Ramond, P. Bouvarel, et J. Rodriguez. Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas. Technical report, https ://hal.archives-ouvertes.fr/hal-01909708/document, 2018.

C. Coron, C. Calenge, C. Giraud, et R. Julliard. Estimation of species relative abundances and habitat preferences using opportunistic data. *arXiv preprint arXiv :1706.08281*, 2017.

D. D. Cox et R. L. Savoy. Functional magnetic resonance imaging (fMRI)"brain reading" : detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19(2) :261–270, 2003.

D. Dacunha-Castelle et E. Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM : Probability and Statistics*, 1 :285–317, 1997.

K. Dang-Van. Macro-micro approach in high-cycle multiaxial fatigue. Dans *Advances in multiaxial fatigue*. ASTM International, 1993.

J.-J. Daudin, F. Picard, et S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18 :173–183, 2008.

S. Dehaene, G. Le Clec'H, L. Cohen, J.-B. Poline, P.-F. van de Moortele, et D. Le Bihan. Inferring behavior from functional brain images. *Nature neuroscience*, 1(7) :549, 1998.

A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

A. Desrochers et A. Clément. A dimensioning and tolerancing assistance model for cad/cam systems. *The International Journal of Advanced Manufacturing Technology*, 9 (6) :352–361, 1994.

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.

S. Evans, P. C. Waller, et 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.

J. Felsenstein. Evolutionary trees from dna sequences : a maximum likelihood approach. *Journal of molecular evolution*, 17(6) :368–376, 1981.

G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache, et J.-B. Poline. Improved detection sensitivity in functional MRI data using a brain parcelling technique. Dans *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 467–474. Springer, 2002.

M. Fop et T. B. Murphy. Variable selection methods for model-based clustering. *Statistics Surveys*, 12 :18–65, 2018.

R. Fouchereau, G. Celeux, et P. Pamphile. Probabilistic modeling of s–n curves. *International Journal of Fatigue*, 68 :217–223, 2014.

J. Friedman, T. Hastie, et R. Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001.

K. Friston, C. Chu, J. Mourão-Miranda, O. Hulme, G. Rees, W. Penny, et J. Ashburner. Bayesian decoding of brain images. *Neuroimage*, 39(1) :181–205, 2008.

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, et R. S. Frackowiak. Statistical parametric maps in functional imaging : a general linear approach. *Human brain mapping*, 2(4) :189–210, 1994.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

P. Gaillard et Y. Goude. Forecasting electricity consumption by aggregating experts ; how to design a good set of experts. Dans *Modeling and stochastic learning for forecasting in high dimensions*, pages 95–115. Springer, 2015.

E. Gassiat et C. Keribin. The likelihood ratio test for the number of components in a mixture with markov regime. *ESAIM : Probability and Statistics*, 4 :25–52, 2000.

E. Gassiat et R. Van Handel. Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory*, 59(2) :1115–1128, 2013.

S. Ghebreab, A. Smeulders, et P. Adriaans. Predicting brain states from fMRI data : Incremental functional principal component regression. Dans *Advances in Neural Information Processing Systems*, pages 537–544, 2008.

C. Giraud, C. Calenge, C. Coron, et R. Julliard. Capitalising on opportunistic data for monitoring species relative abundances. *arXiv preprint arXiv :1407.2432*, 2014.

G. Govaert. Algorithme de classification d'un tableau de contingence. Dans *First international symposium on data analysis and informatics*, pages 487–500, 1977.

G. Govaert. Simultaneous clustering of rows and columns. *Control and Cybernetics*, 24 : 437–458, 1995.

G. Govaert et M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36 (2) :463–473, 2003.

G. Govaert et M. Nadif. Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6) :3233–3245, 2008.

G. Govaert et M. Nadif. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3) :416–425, 2010.

G. Govaert et M. Nadif. *Co-clustering*. John Wiley & Sons, 2013.

A. Gunawardana et W. Byrne. Convergence theorems for generalized alternating minimization procedures. *Journal of machine learning research*, 6(Dec) :2049–2073, 2005.

M. Gyllenberg, T. Koski, E. Reilink, et M. Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2) :542–548, 1994.

R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, et N. H. Shah. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clinical Pharmacology & Therapeutics*, 93(6) :539–546, 2013.

C. Hennig, M. Meila, F. Murtagh, et R. Rocci (eds). *Handbook of cluster analysis*. CRC Press, 2015.

C. Hernandez, C. Keribin, P. Drobinski, et S. Turquety. Statistical modelling of wildfire size and intensity : a step toward meteorological forecasting of summer extreme fire risk. *Annales Geophysicae*, 33(12) :1495–1506, 2015. URL `https://www.ann-geophys.net/33/1495/2015/angeo-33-1495-2015.pdf`.

L. Hubert et P. Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.

J. Jacques et C. Biernacki. Model-based co-clustering for ordinal data. *Computational Statistics & Data Analysis*, 123 :101–115, 2018.

C. Keribin. Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 326(2) :243–248, 1998.

C. Keribin. *Tests de modèles par maximum de vraisemblance*. PhD thesis, Evry-Val d'Essonne, 1999.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

C. Keribin. Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie. *Journal de la Société Française de Statistique*, 151(2) :107–131, 2010. URL `http://journal-sfds.math.cnrs.fr/index.php/J-SFdS/article/view/51`.

C. Keribin. Choix de modèles quand la vraisemblance est incalculable. Dans *47èmes Journées de Statistique de la SFdS*, 2015.

C. Keribin. A note on BIC and the slope heuristic. *Journal de la SFdS*, 160(3) :136–139, 2019. URL `http://journal-sfds.fr/article/view/755/802`.

C. Keribin et D. Haughton. Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families. *Communications in Statistics-Theory and Methods*, 32(7) :1373–1404, 2003.

C. Keribin, G. Govaert, et G. Celeux. Estimation d'un modèle à blocs latents par l'algorithme SEM. Dans *42èmes Journées de Statistique*, 2010.

C. Keribin, V. Brault, G. Celeux, G. Govaert, et al. Model selection for the binary latent block model. Dans *Proceedings of COMPSTAT*, volume 2012, 2012.

C. Keribin, V. Brault, G. Celeux, et G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6) :1201–1216, 2015. URL `https://link.springer.com/article/10.1007/s11222-014-9472-2`.

C. Keribin, G. Celeux, et V. Robert. The latent block model : a useful model for high dimensional data. Dans *61st ISI World Statistics Congress, ISI2017, Marrakech*, 2017.

C. Keribin, Y. Liu, T. Popova, et Y. Rozenholc. A mixture model to characterize genomic alterations of tumors. *Journal de la SFdS*, 160(1) :130–148, 2019. URL `http://journal-sfds.fr/article/view/731/777`.

N. Kriegeskorte, R. Goebel, et P. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10) :3863–3868, 2006.

C. Laclau et V. Brault. Noise-free latent block model for high dimensional data. *Data Mining and Knowledge Discovery*, 33(2) :446–473, 2019.

P. Latouche et S. Robin. Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a w-graph model. Technical report, Technical report, 2013.

P. Latouche, E. Birmelé, C. Ambroise, et al. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1) : 309–336, 2011.

D. D. Lee et H. S. Seung. Algorithms for non-negative matrix factorization. Dans *Advances in neural information processing systems*, pages 556–562, 2001.

B. Li et S. C. Hoi. Online portfolio selection : A survey. *ACM Computing Surveys (CSUR)*, 46(3) :35, 2014.

B. G. Lindsay. Mixture models : theory, geometry and applications. Dans *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

Y. Liu, C. Keribin, T. Popova, et Y. Rozenholc. Statistical estimation of genomic tumoral alterations. Dans *47èmes Journées de Statistique de la SFdS*, 2015.

A. Lomet. *Sélection de modèles pour la classification de données continues*. PhD thesis, Université Technologique de Compiègne, 2012.

B. Long, Z. M. Zhang, et P. S. Yu. Co-clustering by block value decomposition. Dans *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 635–640. ACM, 2005.

M. Marbac, P. Tubert-Bitter, et M. Sedki. Bayesian model selection in logistic regression for the detection of adverse drug reactions. *Biometrical Journal*, 58(6) :1376–1389, 2016.

M. Mariadassou et C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1) :537–573, 2015.

M. Mariadassou, V. Brault, et C. Keribin. Normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de blocs latents. Dans *48èmes Journées de Statistique de la SFdS*, 2016.

P. Massart. *Concentration inequalities and model selection.* Springer, 2007.

L. Massoulié. Community detection thresholds and the weak Ramanujan property. Dans *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

C. Maugis, G. Celeux, et M.-L. Martin-Magniette. Variable selection in model-based clustering : A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11) :3872–3882, 2009.

G. McLachlan et T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

G. McLachlan et D. Peel. *Finite Mixture Models.* John Wiley & Sons Hoboken, NJ, 2000.

G. J. McLachlan et K. E. Basford. *Mixture models : Inference and applications to clustering*, volume 84. M. Dekker New York, 1988.

V. Michel. *Understanding the visual cortex by using classification techniques.* PhD thesis, Université Paris Sud, 2010.

V. Michel, C. Damon, et B. Thirion. Mutual information-based feature selection enhances fMRI brain activity classification. Dans *2008 5th IEEE international symposium on biomedical imaging : From nano to macro*, pages 592–595. IEEE, 2008.

V. Michel, E. Eger, C. Keribin, et B. Thirion. Adaptive multi-class Bayesian sparse regression-an application to brain activity classification. Dans *MICCAI 2009 : fMRI data analysis workshop-Medical Image Computing and Computer Aided Intervention*, page 1, 2009.

V. Michel, E. Eger, C. Keribin, J.-B. Poline, et B. Thirion. A supervised clustering approach for extracting predictive information from brain activation images. Dans *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 7–14. IEEE, 2010a.

V. Michel, E. Eger, C. Keribin, et B. Thirion. Multi-class sparse Bayesian regression for neuroimaging data analysis. Dans *International Workshop on Machine Learning in Medical Imaging*, pages 50–57. Springer, 2010b.

V. Michel, E. Eger, C. Keribin, et B. Thirion. Multiclass sparse Bayesian regression for fMRI-based prediction. *Journal of Biomedical Imaging*, 2011 :2, 2011. URL `http://www.hindawi.com/journals/ijbi/2011/350838/`.

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, et B. Thirion. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45 (6) :2041–2049, 2012. URL `https://arxiv.org/pdf/1104.5304.pdf`.

G. W. Milligan. Clustering validation : results and implications for applied analyses. Dans *Clustering and classification*, pages 341–375. World Scientific, 1996.

M. Palatucci et T. Mitchell. Classification in very high dimensional problems with handfuls of examples. Dans *PKDD*, pages 212–223, 2007.

T. Popova, E. Manié, D. Stoppa-Lyonnet, G. Rigaill, E. Barillot, M. H. Stern, et al. Genome alteration print (GAP) : a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol*, 10(11) :R128–R128, 2009.

R. Rastelli. Exact ICL maximisation for the stochastic block transition model. *Journal de la SFdS*, 160(1) :35–56, 2019.

C. Robert. *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

V. Robert. *Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance*. PhD thesis, Université Paris Saclay, 2017.

V. Robert, G. Celeux, et C. Keribin. Un modele statistique pour la pharmacovigilance. Dans *47emes Journées de Statistique de la SFdS*, 2015.

V. Robert, G. Celeux, C. Keribin, et P. Tubert-Bitter. Modele des blocs latents et sélection de modeles en pharmacovigilance. Dans *48 èmes Journées de Statistique de la SFdS*, 2016.

S. Ryali, K. Supekar, D. A. Abrams, et V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2) :752–764, 2010.

P. B. Ryan, M. J. Schuemie, E. Welebob, J. Duke, S. Valentine, et A. G. Hartzema. Defining a reference set to support methodological research in drug safety. *Drug safety*, 36(1) :33–47, 2013.

A. Samé, C. Ambroise, et G. Govaert. A classification EM algorithm for binned data. *Computational statistics & data analysis*, 51(2) :466–480, 2006.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) : 461–464, 1978.

M. Selosse, J. Jacques, et C. Biernacki. Model-based co-clustering for mixed type data. Technical report, https ://hal.archives-ouvertes.fr/hal-01893457, 2018.

H. Shan et A. Banerjee. Bayesian co-clustering. Dans *2008 Eighth IEEE International Conference on Data Mining*, pages 530–539. IEEE, 2008.

Y. B. Slimen, S. Allio, et J. Jacques. Model-based co-clustering for functional data. *Neurocomputing*, 291 :97–108, 2018.

G. Stoltz et G. Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59(1-2) :125–159, 2005.

H. Teicher et al. On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1) :55–73, 1960.

B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, et J.-B. Poline. Dealing with the shortcomings of spatial normalization : Multi-subject parcellation of fMRI datasets. *Human brain mapping*, 27(8) :678–693, 2006.

B. Thyreau, B. Thirion, G. Flandin, et J.-B. Poline. Anatomo-functional description of the brain : a probabilistic approach. Dans *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

M. E. Tipping. The relevance vector machine. Dans *Advances in neural information processing systems*, pages 652–658, 2000.

T. Tokuda, J. Yoshimoto, Y. Shimizu, K. Yoshida, S. Toki, G. Okada, M. Takamura, T. Yamamoto, S. Yoshimura, Y. Okamoto, et al. Bayesian multiple and co-clustering methods : Application to fMRi data. *IPSJ SIG Notes*, 176(28) :1–5, 2014.

E. P. van Puijenbroek, A. Bate, H. G. Leufkens, M. Lindquist, R. Orre, et 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.

G. Varoquaux, A. Gramfort, J.-B. Poline, et B. Thirion. Brain covariance selection : better individual functional connectivity models using population prior. Dans *Advances in neural information processing systems*, pages 2334–2342, 2010.

Y. R. Wang, P. J. Bickel, et al. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2) :500–528, 2017.

J. Wu. *Model-based clustering and model selection for binned data.* PhD thesis, Supélec, 2014.

J. Wyse et N. Friel. Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2) :415–428, 2012.

J. Yoo et S. Choi. Orthogonal nonnegative matrix tri-factorization for co-clustering : Multiplicative updates on stiefel manifolds. *Information processing & management*, 46 (5) :559–570, 2010.

G. Youness et G. Saporta. Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée*, 52(1) :97–120, 2004.