Contribution à l'analyse Bayésienne de modèles à variables latentes

Pierre Latouche

Université Paris 1 Panthéon-Sorbonne
Laboratoire : Statistique, Analyse, Modélisation Multidisciplinaire, EA 4543

Mémoire d'habilitation à diriger des recherches

Spécialité : Mathématiques appliquées

Pierre LATOUCHE

Contribution à l'analyse Bayésienne de modèles à variables latentes

Date de soutenance : 27/11/2017

Après avis des rapporteurs :
Adrian RAFTERY (Université de Washigton, Stanford)
Judith ROUSSEAU (Université d'Oxford, Paris-Dauphine)
Fabrice ROSSI (Université Paris 1 Panthéon-Sorbonne)

Soutenue publiquement devant le jury composé de :

François CARON	Université d'Oxford	Examinateur
Marie COTTRELL	Université Paris 1 Panthéon-Sorbonne	Examinatrice
Catherine MATIAS	Université Pierre-et-Marie Curie, CNRS	Examinatrice
Adrian RAFTERY	Université de Washington, Stanford	Rapporteur
Christian ROBERT	Université Paris-Dauphine, Warwick	Examinateur
Fabrice ROSSI	Université Paris 1 Panthéon-Sorbonne	Directeur de recherche
Judith ROUSSEAU	Université d'Oxford, Paris-Dauphine	Rapporteur

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Laboratoire SAMM
Centre Pierre Mendès France
Université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac
75634 PARIS Cedex 13

Remerciements

Ce travail est le résultat de collaborations et d'interactions avec de nombreuses personnes que je souhaite vivement remercier.

Je remercie, tout d'abord, Fabrice Rossi d'avoir accepté d'être mon directeur de recherche dans l'écriture de ce mémoire. Savant parmi les savants, il est rare de ne pas apprendre quelque chose en discutant avec lui. J'ai fait le choix d'écrire un document assez complet donc dense et il a pris le temps de le relire en détails. Ses remarques m'ont été d'une aide précieuse. Je remercie également Judith Rousseau et Adrian Raftery d'avoir accepté d'être rapporteurs de mon HDR. Vos conseils m'ont beaucoup apporté. Je regrette déjà le départ de Judith pour Oxford, les discussions dans le train vont me manquer. Je remercie ensuite François Caron et Christian Robert dont je suis les avancées scientifiques avec grand intérêt. Que soit également sincèrement remerciée Marie Cottrell. En tant que directrice puis simplement en tant que collègue, elle a facilité la mise en place de beaucoup de choses dans le laboratoire. Enfin, je remercie chaleureusement Catherine Matias qui a toujours été de bons conseils. Les membres de ce jury font tous référence dans leurs disciplines respectives et je suis très heureux de les rassembler à l'occasion de mon HDR. Ce sont également des personnalités tout à fait accessibles.

Le laboratoire SAMM m'a fourni un cadre de recherche parfait. Les collaborations et les échanges m'ont beaucoup apporté. Je remercie en particulier Jean-Marc Bardet ainsi qu'Annie Millet, Bruno Nazaret, Madalina Olteanu, Julien Random-Furling, et Joseph Rynckiewz.

Je veux ensuite exprimer ma gratitude à Ian Nabney, Christophe Ambroise, et Etienne Birmelé qui ont été les premiers à m'initier à la recherche, en master puis en doctorat. Que soient également remerciés mes co-auteurs Laurent Bergé, Julien Chiquet, Etienne Côme, Nial Friel, Stéphane Lamassé, Sarah Ouadah, Riccardo Rastelli, et Jason Wyse. Je remercie tout particulièrement Stéphane Robin et Charles Bouveyron. J'ai beaucoup appris à vos côtés. Stéphane, ton cerveau fonctionne réellement très vite et bien. J'admire ta capacité d'abstraction et d'imagination. Avant chacune de nos réunions, sache qu'il m'est nécessaire de me doper à la caféine. Charles, brillant chercheur et avant tout mon ami : que de chemin parcouru depuis notre rencontre en 2011. Je repense à nos premières discussions autour du modèle RSM et je vois aujourd'hui l'aboutissement du projet linkage. J'ai eu beaucoup de plaisir à travailler avec toi ces années sur nos multiples projets. L'aventure continue avec notament le projet Topix.

J'adresse également mes remerciements à mes étudiants, notamment Rawya Zreik, Marco Corneli, et Pierre-Alexandre Mattei que j'ai eu la chance d'encadrer en thèse. Vous voir avancer dans vos sujets de recherche a été une grande source de satisfaction. J'ai beaucoup apprécié travaillller avec vous.

J'achève ici l'écriture de ce mémoire. En tapant ces toutes dernières lignes, j'ai une pensée pour mes amis d'enfance avec qui je suis resté très proche. Je pense à Alexis, Renaud, Yannick, et Pierre. Je ne vous lâcherai pas. Je pense à mes parents Serge et Evelyne, ainsi

qu'à mon frère Paul et à ma soeur Alice. Je vous remercie pour votre soutien et votre présence. Paul et Alice, quoi vous que vous entrepreniez, vous allez y arriver. Je remercie également ma belle famille pour sa gentillesse et les moments passés ensemble. Je pense surtout à ma femme Julie et à nos deux merveilleux garçons Arthur et Louis. Je réalise la chance incroyable que j'ai de vous avoir. Il y a mon travail, mes théorèmes et mes algorithmes pour modéliser le monde. Au-dela de ces symboles et de ces chiffres, il y a l'essentiel, c'est-à-dire vous, ma famille. Merci Julie d'avoir été à mes côtés toutes ces années. Merci pour ton amour. Merci aussi pour ton soutien. Et vous, mes fils, soyez remerciés. Je suis très fier d'être votre père.

À Julie, Arthur, et Louis

Table des matières

1 Introduction 1

2 Des réseaux, des textes, des blocs. Modèles statiques 7
- 2.1 Introduction 8
- 2.2 Inférence dans les modèles à blocs stochastiques 9
- 2.3 Recherche de clusters chevauchants 15
- 2.4 Caractérisation des sous-graphes 23
- 2.5 Analyse conjointe de réseaux et de textes 26

3 Des réseaux, des textes, des blocs. Modèles dynamiques 33
- 3.1 Introduction 34
- 3.2 Temps discret 34
- 3.3 Temps continu et segmentation 41

4 De l'étude des graphons 49
- 4.1 Introduction 50
- 4.2 Estimation dans le modèle W-graphe 51
- 4.3 Qualité de l'ajustement : approche Bayésienne 55
- 4.4 Qualité de l'ajustement : tests 61

5 La grande dimension 67
- 5.1 Introduction 68
- 5.2 Régression linéaire parcimonieuse Bayésienne 68
- 5.3 ACP globalement parcimonieuse 72
- 5.4 Sélection de la dimension en ACP Bayésienne 78

6 Conclusion et perspectives 84

1

Introduction

Ce manuscrit rend compte de mes activités de recherche depuis ma thèse de doctorat. Mes travaux s'inscrivent dans le cadre de l'apprentissage statistique et peuvent être organisés à travers quatre axes :

- réseaux, textes, et blocs. Modèles statiques ;
- réseaux, textes, et blocs. Modèles dynamiques ;
- étude des graphons ;
- la grande dimension.

Les chapitres 2 à 5 décrivent mes travaux publiés, un chapitre étant considéré pour chaque axe. Le chapitre 2 introduit des éléments importants pour la compréhension des chapitres 3 et 4. Il décrit notamment le modèle à blocs stochastiques. Le chapitre 5 peut être lu indépendamment. Alors que les chapitres 2 à 4 s'inscrivent exclusivement dans un cadre d'apprentissage statistique non supervisé, sans variable cible, le chapitre 5 s'intéresse également à des éléments d'apprentissage statistique supervisé. Un dénominateur commun à tous ces chapitres est l'utilisation quasi systématique de la théorie Bayésienne pour la caractérisation de l'incertitude des paramètres.

Le chapitre 2 introduit des outils pour l'analyse des réseaux et des textes. L'objectif est à chaque fois de développer une méthodologie statistique, c'est-à-dire un modèle ainsi qu'une procédure d'inférence associée pour l'estimation des paramètres, afin d'extraire des informations pertinentes pour le praticien. Les données considérées sont statiques, autrement dit elles n'évoluent pas au cours du temps. Le chapitre 3 traite des mêmes questions mais cette fois ci pour des données dynamiques. Les modèles doivent être adaptés et nous avons en particulier recours à des processus. Ensuite, le chapitre 4 s'intéresse plus précisément au modèle de W-graphe, basé sur une fonction appelée graphon, généralisant la plupart des outils utilisés pour la modélisation des réseaux. Les techniques d'optimisation et de tests discutées rendent possible l'utilisation de ce modèle pour l'étude de réseaux réels. Enfin, le

chapitre 5 traite de l'étude de la grande dimension où les données ont un nombre conséquent de variables. Il y est notamment fait mention de méthodes pour l'estimation de la dimension intrinsèque des données ainsi que d'approches Bayésiennes alternatives au Lasso pour la régression parcimonieuse.

Tout au long de ce manuscrit, nous donnons des éléments bibliographiques afin d'orienter le lecteur et de situer nos travaux dans la littérature. Les méthodes décrites étant toutes basées sur un modèle génératif, où les données à analyser sont supposées être le résultat de procédés aléatoires, seuls les éléments bibliographiques s'inscrivant dans ce cadre sont essentiellement donnés. Les méthodes heuristiques basées sur le critère de modularité de Newman ne sont par exemple que très peu discutées dans le chapitre 2.

Mes travaux ont donné lieu à ce jour à 15 publications dans des journaux internationaux. Une partie de ces travaux a été réalisée dans le cadre de trois thèses que j'ai co-encadrées : la thèse de Rawya Zreik, soutenue en novembre 2016, ainsi que celles de Marco Corneli et Pierre Alexandre Mattei, dont les soutenances sont prévues pour octobre 2017. La liste de mes publications est donnée ci-dessous et est utilisée pour référencer mes travaux dans le manuscrit. Les articles sont numérotés par ordre chronologique de 1 à 21.

Articles de journaux avec comité de lecture

[1] P. Latouche, E. Birmelé et C. Ambroise. « Overlapping stochastic block models with application to the French political blogosphere ». In : *Annals of Applied Statistics* 5.1 (2011), p. 309–336.

[2] P. Latouche, E. Birmele et C. Ambroise. « Variational Bayesian inference and complexity control for stochastic block models ». In : *Statistical Modelling* 12.1 (2012), p. 93–115.

[3] Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou et S. Lamassé. « The random subgraph model for the analysis of an ecclesiastical network in Merovingian Gaul ». In : *Annals of Applied Statistics* 8.1 (2014), p. 377–405.

[4] P. Latouche, E. Birmelé et C. Ambroise. « Model selection in overlapping stochastic block models ». In : *Electronic Journal of Statistics* 8.1 (2014), p. 762–794.

[5] E. Côme et P. Latouche. « Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood ». In : *Statistical Modelling* 15.6 (2015), p. 564–589.

[6] R. Zreik, P. Latouche et C. Bouveyron. « Classification automatique de réseaux dynamiques avec sous-graphes : étude du scandale Enron ». In : *Journal de la Société Française de Statistique* 156.3 (2015), p. 166–191.

[7] C. Bouveyron, P. Latouche et R. Zreik. « The stochastic topic block model for the clustering of vertices in networks with textual edges ». In : *Statistics and Computing* (2016), p. 1–21.

[8] M. Corneli, P. Latouche et F. Rossi. « Block modelling in dynamic networks with non-homogeneous Poisson processes and exact ICL ». In : *Social Network Analysis and Mining* 6.1 (2016), p. 55–85.

[9] M. Corneli, P. Latouche et F. Rossi. « Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks ». In : *Neurocomputing* 192 (2016), p. 81–91.

[10] P. Latouche, P-A. Mattei, C. Bouveyron et J. Chiquet. « Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression ». In : *Journal of Multivariate Analysis* 146 (2016), p. 177–190.

[11] P. Latouche et S. Robin. « Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models ». In : *Statistics and Computing* 26.6 (2016), p. 1173–1185.

[12] R. Zreik, P. Latouche et C. Bouveyron. « The dynamic random subgraph model for the clustering of evolving networks ». In : *Computational Statistics* (2016), p. 1–33.

[13] J. Wyse, N. Friel et P. Latouche. « Inferring structure in bipartite networks using the latent blockmodel and exact ICL ». In : *Network Science* 5.1 (2017), p. 45–69.

[14] M. Corneli, P. Latouche et F. Rossi. « Multiple change point detection and clustering in dynamic networks ». In : *Statistics and Computing* (à paraître).

[15] P. Latouche, S. Robin et S. Ouadah. « Goodness of fit of logistic regression models for random graphs ». In : *Journal of Computational and Graphical Statistics* (à paraître).

Chapitres de livres

[16] P. Latouche, E. Birmelé et C. Ambroise. « Bayesian methods for graph clustering ». In : *Advances in Data Handling and Business Intelligence*. Springer, 2009.

[17] P. Latouche, E. Birmelé et C. Ambroise. « Overlapping clustering methods for networks ». In : *Handbook of Mixed Membership Models and Their Applications*. Chapman et Hall/CRC, 2014.

[18] R. Zreik, P. Latouche et C. Bouveyron. « Cluster Identification in Maritime Flows with Stochastic Methods ». In : *Maritime Networks : Spatial Structures and Time Dynamics*. Routledge, 2015.

Prépublications

[19] C. Bouveyron, P. Latouche et P-A. Mattei. « Bayesian Variable Selection for Globally Sparse Probabilistic PCA ». 2017.

[20] C. Bouveyron, P. Latouche et P-A. Mattei. « Exact dimensionality selection for Bayesian PCA ». 2017.

[21] S. Ouadah, S. Robin et P. Latouche. « A degree-based goodness-of-fit test for heterogeneous random graph models ». 2017.

Brevet, applications, et logiciels

Mes travaux s'inscrivant en statistique computationnelle ont été essentiellement motivés par des questions soulevées dans le cadre d'applications sur données réelles. Nous donnons ci-dessous la liste des applications pour lesquelles nos développements méthodologiques ont fourni des résultats. En pratique, les développements se sont accompagnés de l'écriture de logiciels et packages. Nous les présentons également ci-dessous. Finalement, un brevet a été déposé.

Brevet

- la méthodologie pour le modèle STBM, implémentée dans la plate-forme web linkage.fr, a donné lieu à un dépôt de brevet.

Études de cas

- réseau métabolique d'Escherichia coli ;
- réseau de transcription de Saccharomyces cerevisiae ;
- étude du cancer du sein / sélection de gènes ;
- réseau d'illustrations et de bandes dessinées ;
- blogosphere politique française ;
- réseau ecclésiastique en Gaule mérovingienne ;
- emails de l'entreprise Enron ;
- co-publications pour la conférence NIPS ;
- votes au congrès Américain ;
- base de données de films notés par des utilisateurs ;
- trois réseaux maritimes ;
- interactions entre les participants de la conférence ACM Hypertext à Turin ;
- flux de vélos entre les stations Santander de Londres ;
- prédiction de la fréquentation du musée d'Orsay en fonction des flux Vélib à Paris.

J'ai été amené à travailler sur d'autres jeux de données réelles pour des publications. Sont indiqués au dessus seulement les jeux de données pour lesquels nous avons réalisé une étude détaillée.

Développements informatiques

- linkage.fr : plate-forme web (Javascript) s'appuyant sur un code C++. Analyse conjointe de réseaux et de textes portés par les connexions ;
- gof network : package R dont le source est en C++. Analyse de l'ajustement du modèle de régression logistique dans le cas des réseaux ;
- spinyReg : package R. Régression parcimonieuse Bayésienne ;
- OSBM : package R dont le source est en C. Recherche de clusters chevauchants dans les réseaux ;
- rambo : package R. Inférence pour le modèle RSM ;
- mixer : package R dont le source est en C++ / fortran. Inférence pour le modèle SBM.

2

Des réseaux, des textes, des blocs. Modèles statiques

2.1 Introduction 8
2.2 Inférence dans les modèles à blocs stochastiques 9
2.2.1 Approche variationnelle 10
2.2.1.a Approximations 10
2.2.1.b Expérimentations numériques 12
2.2.2 Algorithme glouton 13
2.2.2.a Inférence 14
2.2.2.b Expérimentations numériques 14
2.3 Recherche de clusters chevauchants 15
2.3.1 Identifiabilité 16
2.3.2 Estimation fréquentiste 18
2.3.3 Cadre Bayésien. Sélection de modèles 19
2.3.3.a Dérivation d'une borne inférieure 19
2.3.3.b Inférence 20
2.3.3.c Sélection de modèles 21
2.3.4 Expérimentations numériques 21
2.4 Caractérisation des sous-graphes 23
2.4.1 Modèle 24
2.4.2 Inférence 24
2.4.3 Expérimentations numériques 25
2.5 Analyse conjointe de réseaux et de textes 26
2.5.1 Modélisation de la présence de connexions 26
2.5.2 Modélisation des documents 27
2.5.3 Pivot 28
2.5.4 Inférence 28
2.5.5 Sélection de modèles 29
2.5.6 Expérimentations numériques 29

Ce chapitre est consacré au développement de modèles statiques et de techniques d'inférence associées pour l'analyse de réseaux et de textes. Cet axe de recherche est décomposé en quatre sous-axes développés ci-dessous. Le premier traite du modèle à blocs stochastiques qui est un modèle fondateur en analyse des réseaux. Mes travaux dans ce domaine ont donné lieu à deux articles dans des journaux (Latouche, Birmele et Ambroise, 2012 ; Côme et Latouche, 2015) ainsi qu'à un chapitre de livre (Latouche, Birmelé et Ambroise, 2009). Notons que nous avons adapté Côme et Latouche, 2015 au modèle à blocs latents dans un autre article de journal (Wyse, Friel et Latouche, 2017). Ces deux modèles étant très proches, seuls les travaux dans Côme et Latouche, 2015 sont présentés dans ce manuscrit. Le second sous-axe s'intéresse à la recherche de clusters chevauchants dans les réseaux, où un nœud peut appartenir à plusieurs clusters simultanément, et s'est traduit par deux articles dans des journaux (Latouche, Birmelé et Ambroise, 2011 ; Latouche, Birmelé et Ambroise, 2014a) ainsi qu'à un chapitre de livre (Latouche, Birmelé et Ambroise, 2014b). Ensuite, le troisième axe se concentre sur la modélisation des données lorsque une partition est fournie. Il a donné lieu à un article dans un journal (Jernite, Latouche et al., 2014) et à un chapitre de livre (Zreik, Latouche et Bouveyron, 2015b). Enfin, le dernier axe décrit des travaux permettant de réaliser une étude conjointe d'un réseau et d'un ensemble de textes. Le modèle principale proposé est publié dans un article de journal (Bouveyron, Latouche et Zreik, 2016).

2.1 Introduction

Le terme "réseau" est en réalité synonyme de "graphe". En pratique, dans la littérature, il est plutôt fait mention de réseaux lorsque des applications sur données réelles sont décrites. Le terme de graphe apparaît lui généralement lors de la caractérisation de l'objet mathématique. Ainsi, un graphe est constitué d'un ensemble V de n nœuds et d'un ensemble E de connexions entre ces nœuds. Dans le cas d'un réseau social, les nœuds peuvent représenter des individus et deux individus sont connectés s'ils ont interagi à travers au moins un email par exemple. Il existe de nombreux types de graphe, suivant l'information présente sur les nœuds ainsi que les connexions, et le type de connexions. En particulier, si chaque connexion (i, j), pour $i, j \in V$, est une paire orientée, c'est-à-dire qu'il existe une interaction de i vers j, pas nécessairement réciproque, alors les connexions sont appelées arcs. Le graphe associé est dit orienté. Inversement, si les paires sont non-orientées, on parle d'arêtes et le graphe est non orienté. Techniquement, les méthodes d'inférence statistique doivent s'adapter ponctuellement à ces deux situations. En effet, à titre d'exemple, le produit $\prod_{i,j}^{n}$ sur toutes les paires (i, j) est utilisé dans le cas orienté et est remplacé par le produit $\prod_{i\leq j}^{n}$ lors du passage au non orienté. Certaines contraintes de symétrie des paramètres doivent également être respectées dans les procédures d'optimisation. Afin de décrire les modèles statistiques, nous nous appuierons sur la matrice d'adjacence notée X, de taille $n \times n$, telle que $X_{ij} = 1$ si i et j sont connectés, 0 sinon. Notons de plus, que les paires (i, i), de connexions de nœuds avec eux mêmes, ne sont pas retenues dans la plupart des approches décrites dans la littérature. Dans ce chapitre, comme dans les chapitres 3 et 4, ces paires sont donc écartées et $X_{ij} = 0, \forall i$. Ci-dessous, nous verrons des cas où X est autorisée à prendre ses valeurs dans un ensemble autre que simplement $\{0, 1\}$. Notons finalement que certains travaux seront présentés uniquement pour des graphes non orientés et d'autres uniquement pour des graphes orientés. Les passages d'un cas à un autre ne posent pas de problème particulier. Le modèle probabi-

liste de référence pour l'analyse des réseaux est certainement le modèle à blocs stochastiques, stochastic block model (SBM) en anglais (Wang et Wong, 1987 ; Nowicki et Snijders, 2001). Il constitue en particulier le point de départ de nombreux outils statistiques récents pour le clustering des nœuds. Les stratégies alternatives considèrent généralement le modèle à positions latentes (latent position model, Hoff, Raftery et Handcock, 2002a) ou le modèle exponentiel de graphe aléatoire (exponential random graph model, Hoff, Raftery et Handcock, 2002b).

2.2 Inférence dans les modèles à blocs stochastiques

SBM est un modèle de mélange pour des données de type réseau. Il fait l'hypothèse que chaque nœud i est associé à un vecteur binaire latent tiré à partir d'une loi multinomiale :

$$Z_i|\pi \sim \prod_{q=1}^{Q} \pi_q^{Z_i q},$$

où $\pi = (\pi_q)_q$ est un vecteur de probabilités de taille Q tel que $\sum_{q=1}^{Q} \pi_q = 1$. Tous les Z_i sont iid et par construction $\sum_{q=1}^{Q} Z_{iq} = 1, \forall i$. De plus, en terme de codage, $Z_{iq} = 1$ signifie que i appartient au cluster q, 0 sinon. Dans les modèles de mélange classiques, comme les modèles de mélange Gaussien, un vecteur X_i, caractérisant une observation i, est alors directement généré à partir de Z_i, c'est-à-dire en fonction de son cluster d'appartenance. En terme de modèle graphique, ce conditionnement simple, 1 vers 1, autorise le recours à des méthodes d'inférence comme l'algorithme EM (expectation maximisation en anglais, Dempster, Laird et Rubin, 1977). De plus, la vraisemblance des données s'obtient facilement grâce à l'hypothèse d'indépendance. Cela permet par exemple d'avoir recours à des critères de sélection de modèles de type vraisemblance pénalisée comme BIC (bayesian information criterion en anglais, Schwarz, 1978) ou AIC (Akaike information criterion, Akaike, 1974) pour estimer le nombre Q de clusters présents dans des données. Afin de décrire le processus de génération d'un graphe, c'est-à-dire d'un ensemble de connexions entre des paires de nœuds, le modèle SBM s'appuie lui sur un conditionnement 2 vers 1. Ainsi, chaque X_{ij} est tiré aléatoirement suivant une loi de Bernoulli, conditionnellement à Z_i ainsi que Z_j :

$$X_{ij}|Z_{iq}Z_{jr} = 1, \mu \sim \mathcal{B}(\mu_{qr}).$$

Nous notons $\mu = (\mu_{qr})_{qr}$ la matrice de connectivité de taille $Q \times Q$. De plus, Z désigne la matrice de taille $n \times Q$ composée des Z_{iq}. Ce changement de conditionnement a des conséquences multiples en terme d'inférence. En particulier, la loi de Z, sachant la matrice d'adjacence X et les paramètres, ne se factorise pas. A posteriori, les Z_i sont en effet tous dépendants. L'algorithme EM n'est alors plus utilisable directement. De même, la vraisemblance des données ne se factorise plus et a un coût de calcul prohibitif. Les critères BIC et AIC ne sont donc plus utilisables pour l'estimation de Q. De nombreuses méthodes, aussi bien stochastiques que variationnelles, ont par conséquent été développées ces quinze dernières années pour réaliser l'inférence du modèle SBM à l'aide d'approximations. Il est impossible ici d'en dresser une liste exhaustive. Une description de la plupart de ces travaux est donnée dans Goldenberg, Zheng et al., 2010 et Matias et Robin, 2014. Dans la chronologie de développement de ces outils, l'approche de Daudin, Picard et Robin, 2008 a joué un rôle clé. Ainsi, un algorithme variationnel EM (VEM) est proposé pour réaliser le clustering des nœuds et l'estimation des paramètres. La sélection de modèles s'appuie quant à elle sur le

critère ICL (integrated classification likelihood) de Biernacki, Celeux et Govaert, 2000. Le logiciel *mixer* associé fut un des premiers à permettre l'analyse d'un réseau de plusieurs milliers de nœuds et dizaines de milliers de connexions. Malheureusement, ce critère est basé sur des considérations asymptotiques et n'est pas pertinent pour l'estimation de Q, lorsque la valeur de n est faible ou lorsque des clusters ne contiennent que très peu de nœuds. Pour palier à cette limite, nous avons développé deux solutions décrites ci-dessous. La deuxième permet également de réduire considérablement le coût algorithmique de l'inférence.

2.2.1 Approche variationnelle

Nous présentons dans cette section nos travaux publiés dans Latouche, Birmele et Ambroise, 2012. L'objectif premier est de construire un critère de sélection de modèles pour estimer le nombre de clusters Q dans un réseau, sous un modèle SBM. Un cadre Bayésien est retenu et des lois a priori conjuguées sont utilisées pour caractériser les paramètres. Ainsi, le vecteur π est modélisé par une loi de Dirichlet :

$$p(\pi) = \mathrm{Dir}(\pi; n^0 = (n_1^0, \ldots, n_Q^0)).$$

Le vecteur n^0 est régulièrement défini tel que $n_q^0 = 1/2, \forall q$, dans la littérature. La loi de Dirichlet correspond alors à la loi non informative de Jeffreys (Jeffreys, 1946). Une loi uniforme sur le simplexe est également obtenue en fixant $n_q^0 = 1, \forall q$. Les termes de connectivité μ_{qr} sont quant à eux caractérisés par des lois a priori Beta :

$$p(\mu) = \prod_{q \leq r}^{Q} \mathrm{Beta}(\mu_{qr}; \eta_{qr}^0, \zeta_{qr}^0).$$

La loi Beta étant un cas particulier de loi de Dirichlet, les remarques faites juste au dessus quant au choix des paramètres s'appliquent. Nous considérons dans Latouche, Birmele et Ambroise, 2012 l'analyse de réseaux non orientés où la matrice μ est symétrique. La loi a priori de μ introduite est donc définie ici sur la partie triangulaire supérieure de la matrice.

2.2.1.a Approximations

Comme mentionné précédemment, $p(Z|X, \pi, \mu, Q)$ ne se factorise pas à cause de la structure particulière de conditionnement utilisée dans SBM. De la même manière, la loi a posteriori $p(Z, \pi, \mu|X, Q)$ n'a pas de forme analytique. Dans ce cadre Bayésien, nous proposons de recourir à la vraisemblance marginale $p(X|Q)$ pour estimer Q. A nouveau, cette quantité n'a pas d'écriture explicite et ne peut donc pas être maximisée par rapport à Q. Le cadre variationnel offre des éléments de solution et permet de s'attaquer à ces problèmes simultanément. Ainsi, la log-vraisemblance marginale est décomposée :

$$\log p(X|Q) = \mathcal{L}_Q(r) + \mathrm{KL}(r(\cdot)||p(\cdot|X, Q)),$$

où

$$\mathcal{L}_Q(r) = \sum_Z \int\int r(Z, \pi, \mu) \log\left(\frac{p(X, Z, \pi, \mu|Q)}{r(Z, \pi, \mu)}\right) d\pi d\mu,$$

est une borne inférieure de $\log p(X|Q)$ et

$$\mathrm{KL}(r(\cdot)||p(\cdot|X, Q)) = \sum_Z \int\int r(Z, \pi, \mu) \log\left(\frac{p(Z, \pi, \mu|X, Q)}{r(Z, \pi, \mu)}\right) d\pi d\mu,$$

est la divergence de Kullback-Leibler entre $p(Z, \pi, \mu|X, Q)$ inconnue et $r(Z, \pi, \mu)$. Par construction, maximiser la borne inférieure par rapport à $r(Z, \pi, \mu)$ induit une minimisation de la divergence, $\log p(X|Q)$ ne dépendant pas de $r(Z, \pi, \mu)$. Ainsi, l'optimisation de $\mathcal{L}_Q(r)$ permet à la fois de construire une meilleure approximation de la log-vraisemblance marginale à l'aide de la borne elle même. De plus, la minimisation induite de la divergence autorise la caractérisation de $r(Z, \pi, \mu)$ comme approximation de $p(Z, \pi, \mu|X, Q)$. Notons G une partition des éléments dans (Z, π, μ) en P groupes disjoints $(g_1, \ldots, g_P)$ telle que $r(Z, \pi, \mu) = r(G)$. Pour la factorisation $r(G) = \prod_{j=1}^{P} r(g_j)$, le facteur $r(g_j)$, maximisant la borne inférieure, s'obtient alors directement :

$$\log r(g_j) = \mathrm{E}_{G^{\setminus j}}[\log p(X, G|Q)] + \text{const.}$$

L'espérance est ici prise par rapport à toutes les variables dans G sauf g_j et la constante const est utilisée pour la normalisation de $r(g_j)$. Ce résultat donne naissance à un algorithme VBEM (variational Bayes EM en anglais) où tous les facteurs sont estimés de manière itérée. A convergence de la borne inférieure, $\mathcal{L}_Q(r)$ et $r(G)$ sont utilisées comme approximation de $\log p(X|Q)$ et $p(Z, \pi, \mu|X, Q)$, respectivement. Dans Latouche, Birmele et Ambroise, 2012, nous avons considéré la factorisation $r(Z, \pi, \mu) = \prod_{i=1}^{n} r(Z_i) r(\pi) r(\mu)$. La loi $r(Z_i)$ optimale pour le problème de maximisation est multinomiale et s'écrit :

$$r(Z_i) = \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}}.$$

Le vecteur de probabilités $\tau_i = (\tau_{iq})_q$ vérifie $\sum_{q=1}^{Q} \tau_{iq} = 1$ et τ_{iq} est l'approximation variationnelle de la probabilité a posteriori que le nœud i appartienne au cluster q.

Sélection de modèles L'algorithme VBEM fonctionne à Q fixe et estime $\log p(X|Q)$ et $p(Z, \pi, \mu|X, Q)$. Afin d'estimer Q à partir des données, une grille de valeurs $\{1, \ldots, Q_{\max}\}$ est employée. Pour chaque valeur, l'algorithme est utilisé. Finalement, la valeur Q^* maximisant $\mathcal{L}_Q(r)$ est retenue comme estimateur du nombre de clusters. La borne inférieure $\mathcal{L}_Q(r)$, à convergence, est donc utilisée comme critère de sélection de modèles. Dans Latouche, Birmele et Ambroise, 2012, elle est donnée par la proposition suivante.

Proposition 2.1. *La borne $\mathcal{L}_Q(r)$, à convergence de l'algorithme VBEM pour SBM, est donnée par :*

$$\mathcal{L}_Q(r) = \log\left\{\frac{\Gamma(\sum_{q=1}^{Q} n_q^0)\prod_{q=1}^{Q}\Gamma(n_q)}{\Gamma(\sum_{q=1}^{Q} n_q)\prod_{q=1}^{Q}\Gamma(n_q^0)}\right\} + \sum_{q\leq l}^{Q}\log\left\{\frac{\Gamma(\eta_{ql}^0+\zeta_{ql}^0)\Gamma(\eta_{ql})\Gamma(\zeta_{ql})}{\Gamma(\eta_{ql}+\zeta_{ql})\Gamma(\eta_{ql}^0)\Gamma(\zeta_{ql}^0)}\right\} - \sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}\log(\tau_{iq}),$$

où $\Gamma(\cdot)$ est la fonction gamma et

- $n_q = n_q^0 + \sum_{i=1}^{n}\tau_{iq}, \forall q$;
- $\eta_{q\ell} = \eta_{q\ell}^0 + \sum_{i\neq j}^{n} X_{ij}\tau_{iq}\tau_{j\ell}, \forall q \neq \ell$;
- $\eta_{qq} = \eta_{qq}^0 + \sum_{i<j}^{n} X_{ij}\tau_{iq}\tau_{jq}, \forall q$;
- $\zeta_{q\ell} = \zeta_{q\ell}^0 + \sum_{i\neq j}^{n}(1-X_{ij})\tau_{iq}\tau_{j\ell}, \forall q \neq \ell$;
- $\zeta_{qq} = \zeta_{qq}^0 + \sum_{i<j}^{n}(1-X_{ij})\tau_{iq}\tau_{jq}, \forall q$.

2.2.1.b Expérimentations numériques

Nous nous concentrons sur les expérimentations sur données simulées utilisées dans Latouche, Birmele et Ambroise, 2012 afin d'évaluer le critère de sélection de modèles. Le critère ICL, adapté à SBM dans Daudin, Picard et Robin, 2008, est également employé à titre de comparaison. Ainsi, des réseaux sont générés selon un modèle SBM avec des clusters de mêmes proportions $\pi_q = 1/Q$. Différents types de clusters sont considérés. Seuls les résultats pour une structure avec communautés et un cluster de hubs sont ici présentés. La matrice μ s'écrit alors :

$$\mu = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon & \lambda \\ \epsilon & \lambda & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \lambda & \dots & \dots & \dots & \lambda \end{pmatrix},$$

avec $\lambda > \epsilon$. Ainsi, les $Q-1$ premiers clusters correspondent à des communautés où la probabilité de connexion entre des nœuds d'un même cluster est plus importante qu'entre des nœuds de clusters différents. De plus, le dernier cluster rassemble des nœuds appelés hubs ayant une probabilité λ de se connecter avec tous les nœuds, quels que soient leurs clusters d'appartenance. Les paramètres λ, ϵ, et n sont fixés à 0.9, 0.1, et 50, respectivement. Enfin, pour chaque valeur Q dans $\{3, \dots, 7\}$, 100 réseaux sont simulés. Notre méthodologie est appliquée pour une grille $\{1, \dots, 7\}$ de valeurs de Q et la valeur maximisant le critère de sélection de modèles (IL_{vb} ci-dessous) est retenue. De la même manière, l'algorithme VEM de Daudin, Picard et Robin, 2008 est employé et le critère ICL est utilisé sur la même grille. Les résultats sont donnés dans le tableau 2.1. Ils se dégradent à mesure que le nombre de clusters utilisés dans les simulations augmente. En effet, n étant fixé, le nombre moyen de nœuds dans chaque cluster diminue et il devient plus difficile de les distinguer. Les résultats pour le critère de sélection de modèles proposé sont sensiblement meilleurs que ceux obtenus à l'aide d'ICL. Ainsi, pour $Q = 6$ clusters, $Q^* = 6$ sont estimés dans 70% des cas avec IL_{vb} contre 22% des cas seulement avec ICL.

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	12	**88**	0	0
6	0	0	19	59	**22**	0
7	0	3	29	56	12	**0**

(a) $Q_{vrai} \backslash Q_{ICL}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	2	**98**	0	0
6	0	0	1	29	**70**	0
7	0	0	3	34	45	**18**

(b) $Q_{vrai} \backslash Q_{ILvb}$

Table 2.1 Matrices de confusion du critère de sélection de modèles IL_{vb} de Latouche, Birmele et Ambroise, 2012 et du critère ICL. En ligne : valeurs de Q utilisées pour les simulations. En colonne : valeurs estimées. 100 réseaux sont analysés par ligne.

2.2.2 Algorithme glouton

Les méthodes d'inférence pour SBM ont pour la plus plupart un coût algorithmique en $\mathcal{O}(n^2 Q)$, à Q fixe. Techniquement, cela implique que les logiciels ou packages existants basés sur SBM permettent d'analyser des réseaux ayant quelques milliers de nœuds et des dizaines de milliers de connexions. À titre de comparaison, dans un cadre non probabiliste, certains algorithmes d'optimisation du critère de modularité de Newman, 2006 sont en $\mathcal{O}(m)$ où m désigne le nombre de connexions. Or, comme discuté dans l'introduction du chapitre 4, les réseaux réels sont considérés comme creux, c'est-à-dire que la valeur de m est faible au regard du nombre possible de connexions en n^2. En conséquence, ces techniques permettent d'extraire des clusters dans des réseaux ayant des milliards de connexions. En revanche, elles sont limitées puisqu'elles n'autorisent que la détermination de communautés (voir section précédente). Dans Côme et Latouche, 2015, nous avons donc cherché à développer un cadre d'optimisation pour SBM avec un faible coût algorithmique.

Nous avons tout d'abord considéré le critère ICL pour SBM. D'un point de vue général, comme discuté dans Biernacki, Celeux et Govaert, 2000, ce critère se concentre sur la tache de clustering, contrairement à BIC par exemple qui est dédié à l'estimation de la densité. Il est en effet dépendant de la log-vraisemblance complétée des données $\log p(X, Z|\theta, Q)$ où θ désigne les paramètres du mélange. À noter que dans SBM, où $\theta = (\pi, \mu)$, cette log-vraisemblance est calculable explicitement, contrairement à la log-vraisemblance des données observées $\log p(X|\theta, Q)$. Malheureusement, ICL est construit à partir d'approximations de type Stirling ainsi que Laplace, qui ne tiennent que dans un cadre asymptotique. Comme illustré sur données simulées dans la section 2.2.1.b, ce critère de sélection de modèles offre donc des estimations de qualité moyenne, pour des réseaux de taille limitée. Dans Côme et Latouche, 2015, nous avons montré que les lois a priori conjuguées, introduites en section 2.2.1 de ce mémoire, permettaient d'obtenir une expression exacte pour le critère ICL. Cette dernière permet donc de penser à des stratégies d'optimisation applicables pour des valeurs faibles de n ainsi que dans le cadre asymptotique. Notre méthodologie est ici introduite pour des réseaux orientés.

Proposition 2.2. *La log-vraisemblance intégrée des données complétées du modèle SBM s'écrit de manière explicite pour les lois a priori conjuguées de π et μ introduites en section 2.2.1 :*

$$\begin{aligned} ICL_{ex}(Z, Q) &= \log p(X, Z|Q) \\ &= \sum_{k,\ell}^{Q} \log\left(\frac{\Gamma(\eta^0_{k\ell} + \zeta^0_{k\ell})\Gamma(\eta_{k\ell})\Gamma(\zeta_{k\ell})}{\Gamma(\eta_{k\ell} + \zeta_{k\ell})\Gamma(\eta^0_{k\ell})\Gamma(\zeta^0_{k\ell})}\right) + \log\left(\frac{\Gamma(\sum_{k=1}^{Q} n^0_k)\prod_{k=1}^{Q}\Gamma(n_k)}{\Gamma(\sum_{k=1}^{Q} n_k)\prod_{k=1}^{Q}\Gamma(n^0_k)}\right), \end{aligned}$$

où

- $n_q = n^0_q + \sum_{i=1}^{n} Z_{iq}, \forall q$;
- $\eta_{q\ell} = \eta^0_{q\ell} + \sum_{i\neq j}^{n} X_{ij} Z_{iq} Z_{j\ell}, \forall q, \ell$;
- $\zeta_{q\ell} = \zeta^0_{q\ell} + \sum_{i\neq j}^{n} (1 - X_{ij}) Z_{iq} Z_{j\ell}, \forall q, \ell$.

Notons que $ICL_{ex}(Z, Q)$ ressemble à la borne variationnelle de la proposition 2.1. Point notable, ce critère n'a pas de terme d'entropie et les probabilités τ_{iq} sont remplacées par les variables binaires Z_{iq}.

2.2.2.a Inférence

Pour rappel, la matrice Z de taille $n \times Q$ contient les termes Z_{iq} tels que $Z_{iq} = 1$ si i est dans le cluster q, 0 sinon. Dans la proposition 2.2, nous avons fait dépendre le critère à la fois de Z ainsi que de Q afin de mettre en avant qu'il était utilisable pour des SBM à Q différents. En réalité, dans la preuve associée dans Côme et Latouche, 2015, Q correspond au nombre de colonnes non vides de Z. L'inférence à partir d'$ICL_{ex}(Z, Q)$ est donc un problème d'optimisation en nombres entiers dans Z, où le nombre de colonnes est variable. Cette optimisation permet d'estimer simultanément le nombre de clusters Q et de rechercher la meilleure partition associée au sens de l'ICL. Malheureusement, ce problème est NP dur et nous avons donc eu recours à des heuristiques. Ainsi, en partant d'un nombre Q_{up} élevé et d'une initialisation de Z, des étapes d'échange sont autorisées. Un nœud i peut alors changer de cluster si ce changement induit une augmentation de l'ICL. Si plusieurs changement sont valides, celui induisant la plus grande augmentation est appliqué. Cette étape s'arrête une fois qu'à l'issue d'un passage complet dans l'ensemble des nœuds V du réseau, aucun mouvement n'est retenu. Lors d'un échange, si un nœud est seul dans son cluster d'origine, alors le changement induit la disparition du cluster et Q diminue de 1. Cette étape étant terminée, nous autorisons ensuite des mouvements de fusion entre clusters jusqu'à ce que plus aucune fusion n'entraine une augmentation du critère.

Complexité L'algorithme glouton nécessite le calcul des changements de valeurs d'ICL, lorsqu'un nœud change de cluster et que les autres sont fixes. Or, comme montré dans Côme et Latouche, 2015, ces tests peuvent se faire de manière efficace si bien qu'en moyenne, identifier le meilleur changement de clusters pour un nœud donné, a un coût algorithmique de $\mathcal{O}(l + Q^2)$, où l est le degré moyen dans le réseau. Pour rappel, le degré d'un nœud correspond à son nombre de connexions. Finalement, la procédure partant d'un nombre Q_{up} de clusters, le coût complet de l'algorithme est $\mathcal{O}(n(l + Q_{up}^2) + m)$. En pratique, il est attendu que Q_{up} soit choisi de manière à dominer l, le coût se réduit donc à $\mathcal{O}(nQ_{up}^2 + m)$. En comparaison, à Q fixe, les procédures d'inférence pour SBM ont pour la plupart un coût de $\mathcal{O}(n^2 Q)$. La sélection de modèles nécessite alors un balayage de Q et l'utilisation systématique d'un algorithme d'inférence. Dans ce cas, le coût complet de la procédure d'inférence est donc $\mathcal{O}(n^2 Q_{up}^3)$ pour une recherche entre 1 et Q_{up} clusters dans les données.

2.2.2.b Expérimentations numériques

Nous présentons ici une partie seulement de la longue série d'expériences réalisées sur données simulées dans Côme et Latouche, 2015. L'objectif est de montrer que la réduction du coût algorithmique de la procédure d'inférence proposée n'induit pas une perte de la qualité d'estimation lors de la recherche de clusters. Ainsi, des réseaux à 100 nœuds et $Q = 5$ clusters sont générés selon un modèle SBM. Le paramètre π est tel que $\pi_q = 1/Q$. De plus, la matrice μ est telle que $\mu_{qq} = \beta, \forall q$ et $\mu_{qr} = 0.01, \forall q \neq r$. Les clusters correspondent donc à des communautés. Notons que lorsque β atteint la valeur de 0.01, le modèle n'est pas identifiable. Une grille est alors considérée pour le paramètre de connectivité : $\beta \in \{0.01, \ldots, 0.45\}$. Pour chaque valeur de β, 20 réseaux sont générés et notre procédure, appelée GreedyICL ci-dessous, est utilisée pour la recherche des clusters. En pratique, Q_{up} est fixé à 20. La qualité des clusters identifiés est évaluée à l'aide du critère normalisé d'information mutuelle (Vinh, Epps et Bailey, 2010) et de l'ARI (adjusted rand index, Hubert et Arabie, 1985). Pour comparaison, sont employés sur ces mêmes données, l'algorithme de tirage d'allocations latentes de McDaid, Murphy et al., 2013 (COLSBM), la méthode de sélection de

modèles décrite à la section précédente, implémentée dans le package R mixer, l'algorithme de clustering spectral de Shi et Malik, 2000 (SPECTRAL), et la procédure variationnelle de Hofman et Wiggins, 2008 (VBMOD) dédiée à la recherche de communautés. Les résultats sont donnés dans la figure 2.1. De manière intéressante, GreedyICL, développée à l'origine afin de réduire le coût algorithmique de l'inférence, produit une meilleure estimation des clusters cachés dans les données. La procédure produit en particulier des résultats plus pertinents que COLSBM, méthode de référence pour SBM. Des conclusions similaires sont tirées dans Côme et Latouche, 2015, dans tous les scénarios de simulations considérés.

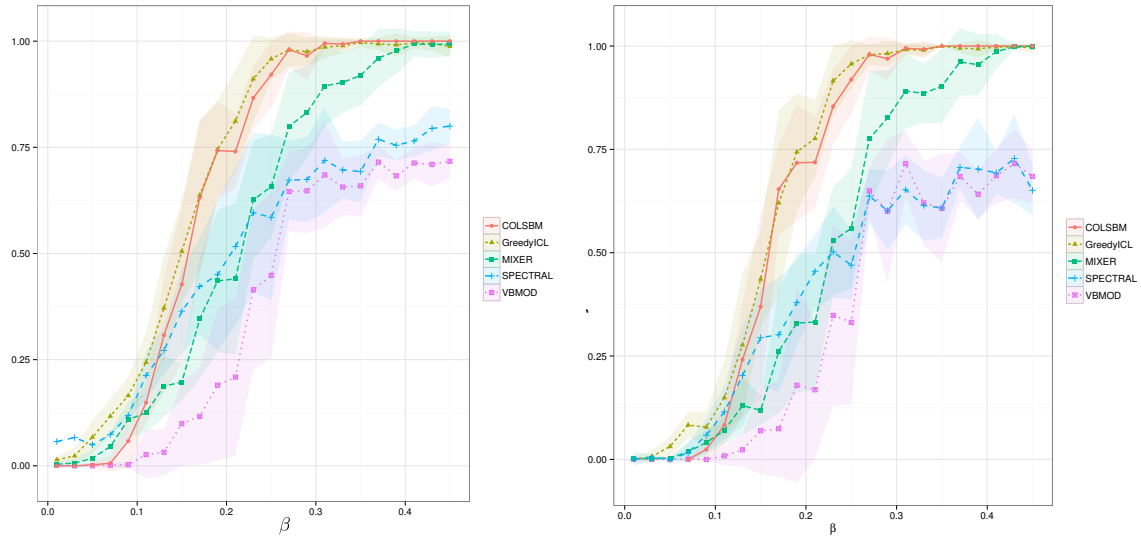


Figure 2.1 – Moyenne du critère normalisé d'information mutuelle (gauche) et de l'ARI ajusté (droite) sur 20 réseaux simulés pour chaque valeur de β dans $\{0.01, \ldots, 0.43\}$.

2.3 Recherche de clusters chevauchants

La plupart des méthodes de clustering de nœuds dans les réseaux construisent des partitions. Autrement dit, chaque nœud d'un réseau se voit associé à un cluster et à un seul. Malheureusement, dans de nombreuses applications réelles, il est attendu des objects d'étude qu'ils appartiennent à plusieurs groupes simultanément. Certaines protéines, appelées *moonlighting proteins*, ont par exemple plusieurs fonctions dans les cellules. Dans un réseau de co-citations entre scientifiques, il est également commun de trouver des chercheurs intervenant dans des disciplines variées. Ranger chaque nœud dans un unique cluster est donc un facteur limitant.

Le travail de Airoldi, Blei et al., 2008 constitue la première véritable solution, basée sur un modèle probabiliste, permettant la recherche de clusters chevauchants. Le modèle associé, appelé mixed membership SBM (MMSB), a deux inconvénients majeurs. Tout d'abord, il utilise un nombre conséquent de paramètres et de vecteurs cachés à estimer. Dans l'article d'origine, son utilisation est donc réservée à des réseaux de quelques dizaines de nœuds. Notons que cette limite est depuis levée grâce à l'algorithme variationnel stochastique de Gopalan et Blei, 2013. De plus, d'un point de vue modèle, la construction des arêtes ou des arcs n'est pas influencée par le fait qu'un nœud appartient à plusieurs clusters. Chaque connexion possible est ainsi caractérisée par des paramètres et vecteurs latentes spécifiques,

et il n'y a pas d'influence entre les clusters.

Nous avons donc proposé un nouveau modèle, appelé overlapping SBM (OSBM), dans Latouche, Birmelé et Ambroise, 2011. Il est ici présenté dans le cas de réseaux orientés. Le tirage aléatoire *iid*, à partir d'une loi multinomiale, des vecteurs cachés caractérisant l'appartenance simple au cluster est remplacée par un tirage *iid* à partir d'un produit de lois de Bernoulli :

$$Z_i|\pi \sim \prod_{q=1}^{Q} \pi_q^{z_{iq}}(1-\pi_q)^{1-z_{iq}}, \forall i.$$

Ainsi, chaque vecteur binaire Z_i peut avoir plusieurs composantes à 1 témoignant ainsi de l'appartenance simultanée de i à plusieurs clusters. Il peut également contenir que des 0. En pratique, cette propriété est utile pour identifier les nœuds n'ayant pas ou très peu de connexions et qui sont de toute manière retirés au moment d'interpréter les résultats du clustering. Contrairement au vecteur de probabilités contrôlant le poids des clusters dans la loi multinomiale, le vecteur π n'est ici pas sujet à des contraintes de type simplexe. Il est donc tout à fait possible d'avoir $\sum_{q=1}^{Q} \pi_q > 1$.

Une fois les clusters construits, les arcs sont tirés aléatoirement selon une loi de Bernoulli conditionnelle :

$$X_{ij}|Z_i, Z_j, a \sim \mathcal{B}\big(X_{ij}; g(a_{Z_i,Z_j})\big), \forall (i \neq j),$$

où

$$a_{Z_i,Z_j} = Z_i^\intercal W Z_j + Z_i^\intercal U + V^\intercal Z_j + W^*.$$

$g(x) = (1+e^{-x})^{-1}$ désigne la fonction logistique, W est une matrice réelle $Q \times Q$, U et V sont des vecteurs réels à Q composantes. Finalement, W^* est un réel permettant de contrôler la présence globale d'arcs, c'est-à-dire la densité du réseau. Notons qu'en définissant $\tilde{Z}_i = (Z_i, 1)^\intercal$, et

$$\tilde{W} = \begin{pmatrix} W & U \\ V^\intercal & W^* \end{pmatrix},$$

le terme de connectivité prend la forme compacte suivante $a_{Z_i,Z_j} = \tilde{Z}_i^\intercal \tilde{W} \tilde{Z}_j$. L'ensemble des paramètres d'OSBM à estimer est donc $(\pi, \tilde{W})$.

2.3.1 Identifiabilité

Avant de dériver une procédure d'estimation d'OSBM sur données réelles, nous avons d'abord étudié l'identifiabilité de ce modèle dans Latouche, Birmelé et Ambroise, 2011. Nos résultats et preuves sont basés sur les travaux fondateurs de Allman, Matias et Rhodes, 2009 qui, à l'aide d'un théorème de Kruskal, 1976 ; Kruskal, 1977 ont caractérisé des conditions d'identifiabilité pour des modèles à variables latentes.

La cadre de démonstration part du constat simple qu'un modèle OSBM à Q clusters peut être réécrit sous la forme d'un modèle SBM à 2^Q clusters. En effet, notant $C, D \in \{0,1\}^Q$ tel que $\sum_{q=1}^{Q} C_q = \sum_{q=1}^{Q} D_q = 1$, considérons la fonction $\phi(\cdot,\cdot)$ telle que :

$$\phi : \quad (\pi, \tilde{W}) \quad \rightarrow \quad (\gamma, \mu) \ ,$$

où $\gamma = (\gamma_C)_C$

$$\gamma_C = \prod_{q=1}^{Q} \pi_q^{C_q}(1-\pi_q)^{1-C_q},$$

et $\mu = (\mu_{C,D})_{C,D}$

$$\mu_{C,D} = g(C^\intercal W D + C^\intercal U + V^\intercal D + W^*).$$

Il est alors évident qu'un modèle OSBM de paramètres $(\pi, \tilde{W})$ induit la même mesure sur l'ensemble des graphes à n nœuds qu'un modèle SBM de paramètres (γ, μ).

Dans leurs travaux, Allman, Matias et Rhodes, 2009 ont montré qu'un modèle SBM est identifiable partout sauf dans un ensemble de paramètres de mesure de Lebesgue nulle. En dehors de cet ensemble critique, les paramètres (γ, μ) et $(\gamma^{'}, \mu^{'})$ engendrent la même mesure si et seulement si $(\gamma^{'}, \mu^{'}) = P_\nu(\gamma, \mu)$ où ν est une permutation sur $\{0,1\}^Q$, $\gamma'_C = \gamma_{\nu(C)}$, et $\mu'_{C,D} = \mu_{\nu(C),\nu(D)}$. Ainsi, étudier l'identifiabilité générale d'OSBM est équivalent à étudier les paramètres vérifiant $\phi(\pi^{'}, \tilde{W}^{'}) = P_\nu(\phi(\pi, \tilde{W}))$ pour une certaine permutation ν sur $\{0,1\}^Q$.

Il est trivial de montrer qu'une permutation simultanée des composantes de π et des lignes/colonnes de $\tilde{W}$ génère des paramètres $(\pi^{'}, \tilde{W}^{'})$ vérifiant cette condition. Cette propriété était attendue car caractéristique des modèles de mélange (McLachlan et Peel, 2004). Nous avons montré qu'il existait une autre famille d'opérations, que nous avons appelées *inversions*, permettant de vérifier cette condition. Nous définissons une A inversion ($A \in \{0,1\}^Q$) de la manière suivante :

$$I_A : \ (\pi, \tilde{W}) \ \rightarrow \ (\pi^{'}, \tilde{W}^{'}) \ ,$$

où

$$\pi^{'}_q = \begin{cases} 1 - \pi_q & \text{si } A_q = 1 \\ \pi_q & \text{sinon} \end{cases} ,$$

et

$$\tilde{W}^{'} = M_A^\intercal \tilde{W} M_A.$$

La matrice M_A est ici donnée par :

$$M_A = \begin{pmatrix} I - 2\text{diag}(A) & A \\ 0 \dots 0 & 1 \end{pmatrix},$$

avec $\text{diag}(A)$ la matrice diagonale $Q \times Q$ dont la diagonale est le vecteur A.

Proposition 2.3. *Pour tout $A \in \{0,1\}^Q$, soit ν la permutation sur $\{0,1\}^Q$ définie par :*

$$\forall C \in \{0,1\}^Q, \nu(C)_i = \begin{cases} 1 - C_i & \textit{si } A_i = 1 \\ C_i & \textit{sinon} \end{cases} .$$

Alors, pour tous paramètres $(\pi, \tilde{W})$ du modèle OSBM, $\phi(I_A(\pi, \tilde{W})) = P_\nu(\phi(\pi, \tilde{W}))$.

Cette proposition a été à la base du développement du théorème suivant :

Théorème 2.1. *Le modèle OSBM est identifiable partout sauf dans un ensemble de paramètres de mesure de Lebesgue nulle. En dehors de cet ensemble, $\phi(\pi, \tilde{W}) = \phi(\pi^{'}, \tilde{W}^{'}) \Leftrightarrow (\pi, \tilde{W}) \sim (\pi^{'}, \tilde{W}^{'})$ où $\sim$ est une relation d'équivalence telle que :*

$$(\pi, \tilde{W}) \sim (\pi^{'}, \tilde{W}^{'}) \quad \textit{if } \exists \sigma, A \quad | \quad (\pi^{'}, \tilde{W}^{'}) = I_A(P_\sigma(\pi, \tilde{W})).$$

2.3.2 Estimation fréquentiste

La vraisemblance du modèle OSBM $p(X|\pi, \tilde{W}, Q) = \sum_Z p(X, Z|\pi, \tilde{W}, Q)$ fait intervenir 2^{nQ} termes et ne peut donc pas être maximisée sous cette forme. Comme pour le modèle SBM non chevauchant, la loi des Z sachant les paramètres et les données X n'a pas de forme analytique et l'algorithme EM ne peut donc pas être utilisé pour estimer les paramètres. À partir de deux approximations variationnelles et des travaux de Jaakkola et Jordan, 2000 sur la régression logistique Bayésienne, nous avons construit une borne inférieure de la log-vraisemblance des données.

Proposition 2.4. *Pour toute matrice ξ positive réelle de taille $n \times n$ et toute distribution $r(Z) = \prod_{i=1}^{n} r(Z_i)$ avec $r(Z_i) = \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}} (1 - \tau_{iq})^{1 - Z_{iq}}$, une borne inférieure de la log-vraisemblance des données est donnée par :*

$$\begin{aligned}
\mathcal{L}(r;\, \pi, \tilde{W}, \xi) = \sum_{i \neq j}^{n} \Big\{ & \big(X_{ij} - \frac{1}{2}\big) \tilde{\tau}_i^\intercal \tilde{W} \tilde{\tau}_j + \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} \\
& - \lambda(\xi_{ij}) \Big(\mathrm{Tr}\Big(\tilde{W}^\intercal \tilde{E}_i \tilde{W} \Sigma_j \Big) + \tilde{\tau}_j^\intercal \tilde{W}^\intercal \tilde{E}_i \tilde{W} \tilde{\tau}_j - \xi_{ij}^2 \Big) \Big\} \\
& + \sum_{i=1}^{n} \sum_{q=1}^{Q} \{ \tau_{iq} \log \pi_q + (1 - \tau_{iq}) \log(1 - \pi_q) \} \\
& - \sum_{i=1}^{n} \sum_{q=1}^{Q} \{ \tau_{iq} \log \tau_{iq} + (1 - \tau_{iq}) \log(1 - \tau_{iq}) \} .
\end{aligned}$$

À l'image de la définition de $\tilde{Z}_i$, $\tilde{\tau}_i = (\tau_i, 1)^\intercal$. Le paramètre τ_{iq} s'interprète comme l'approximation variationnelle de la probabilité, sachant les paramètres $(\pi, \tilde{W})$ et les données dans X, que le nœud i appartienne au cluster q. De plus

$$\Sigma_i = \begin{pmatrix} \tau_{i1}(1 - \tau_{i1}) & 0 & \dots & 0 \\ 0 & \ddots & \dots & \vdots \\ \vdots & 0 & \tau_{iQ}(1 - \tau_{iQ}) & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix},$$

et $\tilde{E}_i = \Sigma_i + \tilde{\tau}_i \tilde{\tau}_i^\intercal$. La fonction $\lambda(\cdot)$ est définie par :

$$\lambda(x) = \frac{1}{4x} \tanh(\frac{x}{2}) = \frac{1}{2x}(g(x) - \frac{1}{2}), \forall x \in \mathbb{R}.$$

Le problème de maximisation de la vraisemblance des données peut donc être remplacé par un problème d'optimisation de $\mathcal{L}$ par rapport à $r(Z)$, π, $\tilde{W}$, et ξ. L'algorithme d'inférence que nous avons proposé dans Latouche, Birmelé et Ambroise, 2011 estime donc ces termes, de manière alternée, jusqu'à convergence de la borne. Notons que la forme paramétrique imposée à $r(Z)$ dans la proposition 2.4 réduit l'optimisation fonctionnelle de $\mathcal{L}$ par rapport à $r(Z)$ à un problème d'optimisation de fonction par rapports aux paramètres τ_{iq}. Les termes ξ_{ij} jouent ici un rôle crucial. Ils permettent de construire des approximations locales de la fonction logistique, pour tous les arcs du réseau, sans lesquelles un algorithme VEM classique n'est pas dérivable.

2.3.3 Cadre Bayésien. Sélection de modèles

L'algorithme d'inférence proposé dans Latouche, Birmelé et Ambroise, 2011 permet de réaliser le clustering des nœuds à partir des probabilités τ_{iq} et l'estimation des paramètres du modèle. En revanche, le cadre fréquentiste considéré dans ce papier n'autorise pas la dérivation d'un critère de sélection de modèles simple afin d'estimer le nombre de clusters présents dans les données. Comme pour le modèle SBM non chevauchant, les critères BIC et AIC ne peuvent par exemple pas être utilisés directement car ils font intervenir la log-vraisemblance des données dont le coût algorithmique d'évaluation est exponentiel. Pour résoudre ce problème, nous avons considéré un cadre Bayésien dans Latouche, Birmelé et Ambroise, 2014a où les paramètres du modèle OSBM se voient caractérisés par des lois a priori.

Pour le vecteur π intervenant dans un produit de loi de Bernoulli, nous avons eu recours à un produit de lois Beta :

$$p(\pi) = \prod_{q=1}^{Q} \text{Beta}(\pi_q; \eta_q^0, \zeta_q^0).$$

Pour une discussion sur le choix des hyperparamètres η_q^0 et ζ_q^0, nous renvoyons le lecteur à la section 2.2.1. Afin de modéliser la matrice réelle $\tilde{W}$ de taille $(Q+1)\times(Q+1)$, l'opérateur vec est utilisé. Ce dernier permet de coller les colonnes d'une matrice sous la forme d'un seul vecteur ligne. Une distribution normale est alors considérée, d'espérance $\tilde{W}_0$ et de matrice de variance covariance $S_0 = I_{(Q+1)^2}/\beta$:

$$p(\tilde{W}^{\text{vec}}|\beta) = \mathcal{N}(\tilde{W}^{\text{vec}};\ \tilde{W}_0^{\text{vec}}, \frac{I_{(Q+1)^2}}{\beta}).$$

Finalement, une loi a priori Gamma est associée au terme d'inverse variance β :

$$p(\beta) = \text{Gam}(\beta;\ a_0, b_0).$$

Par construction, la loi Gamma est informative. Pour limiter son influence sur la loi a posteriori, il est courant dans la littérature de fixer a_0 et b_0 à de faibles valeurs, proches de zéro.

2.3.3.a Dérivation d'une borne inférieure

Dans ce cadre fixé, l'objectif de l'inférence est double. A l'image des travaux réalisés pour le modèle SBM Bayésien et présentés en section 2.2.1, à nombre de clusters Q fixé, nous cherchons à estimer la loi a posteriori des paramètres ainsi que des vecteurs latents Z, sachant les données X. Outre le fait d'obtenir un clustering des nœuds du réseau, cela permet également de mesurer l'incertitude quant à la qualité de l'estimation. Le second objectif est d'approcher la loi marginale des données afin de proposer un estimateur de Q. L'algorithme VBEM permet de répondre à ces deux problématiques simultanément. Malheureusement, il ne peut être dérivé ici directement à cause de la fonction logistique qui ne permet pas d'obtenir certaines espérances conditionnelles de manière explicite. Comme dans Latouche, Birmelé et Ambroise, 2011, nous avons construit deux niveaux d'approximations variationnelles dans Latouche, Birmelé et Ambroise, 2014a afin d'obtenir une borne inférieure de la log-vraisemblance marginale, borne qui peut ensuite être considérée dans un algorithme d'optimisation.

Proposition 2.5. *Pour toute matrice ξ positive réelle de taille $n \times n$ et toute distribution $r(Z, \pi, \tilde{W}, \beta)$, une borne inférieure de la log-vraisemblance marginale est donnée par :*

$$\mathcal{L}_Q(r;\, \xi) = \sum_Z \int\int\int r(Z, \pi, \tilde{W}, \beta) \log \big(\frac{h_Q(Z, \tilde{W}, \xi) p(Z, \pi, \tilde{W}, \beta | Q)}{r(Z, \pi, \tilde{W}, \beta)} \big) d\pi d\tilde{W} d\beta,$$

où

$$\log h_Q(Z, \tilde{W}, \xi) = \sum_{i \neq j}^{n} \left\{ (X_{ij} - \frac{1}{2}) a_{Z_i, Z_j} - \frac{\xi_{ij}}{2} + \log g(\xi_{ij}) - \lambda(\xi_{ij})(a^2_{Z_i, Z_j} - \xi^2_{ij}) \right\},$$

et $\lambda(x) = (g(x) - 1/2)/(2x), \forall x \in \mathbb{R}$, comme précédemment.

2.3.3.b Inférence

Dans l'hypothèse où la matrice ξ est connue, la borne introduite dans la proposition 2.5 peut être maximisée par rapport à $r(Z, \pi, \tilde{W}, \beta)$ en imposant des contraintes sur l'espace fonctionnel d'optimisation. Nous avons ainsi considéré la forme factorisée suivante :

$$r(Z, \pi, \tilde{W}, \beta) = r(\pi) r(\tilde{W}) r(\beta) \prod_{i=1}^{n} \prod_{q=1}^{Q} r(Z_{iq}).$$

Un algorithme VBEM peut alors être dérivé à partir de la borne inférieure $\mathcal{L}_Q(r;\, \xi)$ et les facteurs dans r sont obtenus par optimisation alternée. À r fixée, la borne peut également être maximisée par rapport à la matrice ξ afin d'améliorer les approximations locales de la fonction logistique. Ce cadre donne lieu à une procédure d'optimisation alternée par rapport à r et ξ, jusqu'à convergence de la borne. Ci-dessous, la notation $A_{\cdot q}$ (resp $A_{q\cdot}$) pour une matrice A désigne la colonne (resp ligne) q de la matrice.

Théorème 2.2. *Les facteurs maximisant la borne $\mathcal{L}_Q(r; \xi)$ de la proposition 2.5 sont donnés ci-dessous. Les termes d'espérance sont déterminés par rapport aux facteurs.*

$r(Z_{iq}) = \mathcal{B}(Z_{iq}; \tau_{iq})$ **avec :**

$$\tau_{iq} = g\Big\{ \psi(\eta^N_q) - \psi(\zeta^N_q) + \sum_{j \neq i}^{n} (X_{ij} - \frac{1}{2}) \tilde{\tau}_j^\intercal (\tilde{W}_n^\intercal)_{\cdot q} + \sum_{j \neq i}^{n} (X_{ji} - \frac{1}{2}) \tilde{\tau}_j^\intercal (\tilde{W}_n)_{\cdot q}$$
$$- \mathrm{Tr}\Big(\big(\Sigma^{'}_{qq} + 2 \sum_{l \neq q}^{Q+1} \tilde{\tau}_{il} \Sigma^{'}_{ql}\big) \big(\sum_{j \neq i}^{n} \lambda(\xi_{ij}) \tilde{E}_j \big) + \big(\Sigma_{qq} + 2 \sum_{l \neq q}^{Q+1} \tilde{\tau}_{il} \Sigma_{ql} \big) \big(\sum_{j \neq i}^{n} \lambda(\xi_{ji}) \tilde{E}_j \big) \Big) \Big\},$$

où $\psi(\cdot)$ est la fonction digamma, $\Sigma_{ql} = \mathrm{E}_{\tilde{W}_{\cdot q}, \tilde{W}_{\cdot l}}[\tilde{W}_{\cdot q} \tilde{W}_{\cdot l}^\intercal]$, et $\Sigma^{'}_{ql} = \mathrm{E}_{\tilde{W}_{q\cdot}, \tilde{W}_{l\cdot}}[\tilde{W}_{q\cdot}^\intercal \tilde{W}_{l\cdot}]$.

Les matrices $\tilde{E}_i$ sont définies comme pour l'algorithme VEM décrit précédemment.

$r(\tilde{W}^{\mathrm{vec}}) = \mathcal{N}(\tilde{W}^{\mathrm{vec}}; \tilde{W}^{\mathrm{vec}}_n, S_n)$ **avec :**

$$\tilde{W}^{\mathrm{vec}}_n = S_n \left\{ \sum_{i \neq j}^{n} (X_{ij} - \frac{1}{2}) \tilde{\tau}_j \otimes \tilde{\tau}_i \right\},$$

et

$$S_n^{-1} = \frac{a_n}{b_n} I_{(Q+1)^2} + 2 \sum_{i \neq j}^{n} \lambda(\xi_{ij}) (\tilde{E}_j \otimes \tilde{E}_i),$$

où $\otimes$ désigne le produit de Kronecker.

$r(\pi) = \prod_{q=1}^{Q} \mathrm{Beta}(\pi_q; \eta_q^n, \zeta_q^n)$ **avec :**

$$\eta_q^n = \eta_q^0 + \sum_{i=1}^{n} \tau_{iq} \qquad et \qquad \zeta_q^n = \zeta_q^0 + n - \sum_{i=1}^{n} \tau_{iq}.$$

$r(\beta) = \mathrm{Gam}(\beta; a_n, b_n)$ **avec :**

$$a_n = a_0 + \frac{(Q+1)^2}{2} \qquad et \qquad b_n = b_0 + \frac{1}{2}\mathrm{Tr}(S_n) + \frac{1}{2}(\tilde{W}_n^{\mathrm{vec}})^\intercal \tilde{W}_n^{\mathrm{vec}}.$$

Proposition 2.6. *Un estimateur du paramètre ξ_{ij} de la matrice ξ maximisant la borne $\mathcal{L}_Q(r; \xi)$ de la proposition 2.5 est donné par :*

$$\hat{\xi}_{ij} = \sqrt{\mathrm{Tr}\Big(\big(S_n + \tilde{W}_n^{\mathrm{vec}} (\tilde{W}_n^{\mathrm{vec}})^\intercal\big)(\tilde{E}_j \otimes \tilde{E}_i)\Big)}.$$

2.3.3.c Sélection de modèles

Jusqu'à maintenant, le nombre de clusters Q a été supposé fixe et connu. La procédure d'inférence d'écrite à la section précédente permet alors de maximiser la borne inférieure $\mathcal{L}_Q(r; \xi)$ de la log-vraisemblance marginale, par rapport à r et ξ. Nous avons donc proposé dans Latouche, Birmelé et Ambroise, 2014a de remplacer la log-vraisemblance marginale par son approximation variationnelle. Le principe méthodologique est le même que précédemment. Un ensemble de valeurs de Q est considéré et pour chaque valeur, la valeur de $\mathcal{L}(r; \xi)$, à convergence, est utilisée comme approximation de $\log p(X|Q)$. La valeur Q^* maximisant ce critère est alors retenue comme estimateur du nombre de clusters présents dans les données.

2.3.4 Expérimentations numériques

Nous présentons ici une partie des résultats obtenus à l'aide de simulations dans Latouche, Birmelé et Ambroise, 2011 et Latouche, Birmelé et Ambroise, 2014a. La qualité des clusters identifiés, à Q fixé, est évaluée pour l'algorithme VEM de Latouche, Birmelé et Ambroise, 2011. Notons que des résultats similaires ont été publiés dans Latouche, Birmelé et Ambroise, 2014a dans le cadre Bayésien. La procédure de sélection de modèles est testée à partir des travaux dans Latouche, Birmelé et Ambroise, 2014a.

Nous avons tout d'abord comparé l'algorithme VEM pour OSBM, permettant de chercher des clusters chevauchants dans les données, avec des méthodes de clustering de nœuds. Des précautions doivent être prises pour évaluer ce type de résultats. En effet, en clustering, les critères d'évaluation de la qualité des clusters identifiés, comme l'ARI, ne sont utilisables que sur des partitions, c'est-à-dire sans chevauchement entre les clusters. Nous nous sommes appuyés sur un critère inspiré des travaux de Heller et Ghahramani, 2007 ; Heller, Williamson et Ghahramani, 2008. Ainsi, soient Z la matrice caractérisant les clusters simulés à rechercher et $\hat{Z}$ celle associée à l'estimation d'une méthode. La distance L_2 $d(P, \hat{P})$ entre $P = ZZ^\intercal$ et $\hat{P} = \hat{Z}\hat{Z}^\intercal$ est alors calculée. Les matrices $n \times n$ de la forme $ZZ^\intercal$ permettent de compter le nombre de clusters en commun entre chaque paire de nœuds, et sont invariantes par rapport aux permutations des colonnes de Z.

Les résultats présentés en figure 2.3 ont été obtenus à partir de 100 simulations de réseaux à $n = 100$ nœuds, construits à partir de structures en étoiles et de communautés. Chaque

étoile correspond en réalité à deux clusters et les connexions se font entre des noeuds de clusters différents, à l'inverse des communautés où les connexions se font prioritairement entre des nœuds d'un même cluster. Un exemple de réseau est fourni en figure 2.2. La méthode CFinder (Palla, Derenyi et al., 2006) est capable de chercher des clusters chevauchants dans les réseaux, à l'instar de MMSB. En revanche, contrairement aux autres approches considérées, elle n'est pas basée sur un modèle probabiliste. Le modèle SBM est ici employé comme référence, la méthodologie d'inférence associée n'étant pas en mesure d'identifier des clusters chevauchants. Plus les valeurs de distance $d(P, \hat{P})$ sont faibles, meilleurs sont les résultats de clustering. La figure 2.3 illustre donc clairement la pertinence de l'algorithme VEM pour OSBM, au regards des autres méthodes. Des résultats similaires ont été obtenus à l'aide de réseaux construits uniquement à partir de communautés

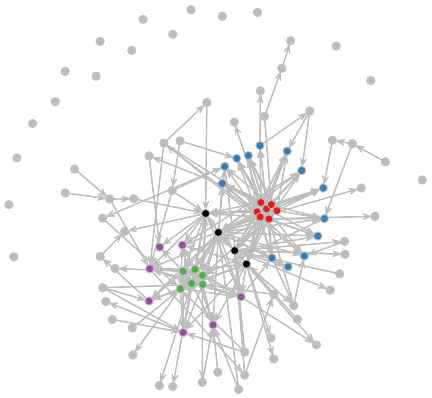


Figure 2.2 – Un exemple de réseau simulé à partir d'étoiles et de communautés. Les clusters en vert et en rouge correspondent à des communautés. Les étoiles, représentées en violet et en bleu, reçoivent des arcs entrants depuis les communautés en vert et en rouge, respectivement. Les nœuds en noir appartiennent aux deux étoiles simultanément. Les nœuds en gris ont un vecteur Z_i dont toutes les composantes sont nulles, et n'appartiennent donc à aucun cluster.

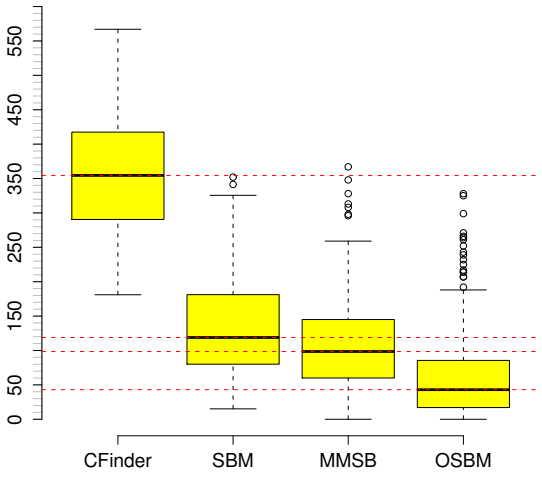


Figure 2.3 – Distance L_2 $d(P, \hat{P})$ pour 100 réseaux simulés à partir d'étoiles et de communautés. Plus les valeurs de distance sont faibles, meilleures sont les estimations associées des clusters chevauchants.

Les résultats d'évaluation du critère de sélection de modèles proposé dans Latouche,

Birmelé et Ambroise, 2014a sont présentés dans le tableau 2.2. Ils correspondent à des réseaux construits à partir de communautés chevauchantes à $n = 100$ nœuds. Ainsi, six scénarios sont considérés en fonction des probabilités de connexion intra cluster et pour des clusters de même taille ou de tailles différentes. Pour chaque nombre de clusters entre 2 et 7, et pour chaque scénario, 100 réseaux sont analysés. Les nombres de clusters estimés sont reportés dans le tableau. Les résultats se dégradent comme attendu à mesure que les probabilités intra cluster diminuent et que les tailles de clusters varient. Néanmoins, le critère dérivé à partir des approximations variationnelles permet une estimation pertinente du nombre de clusters chevauchants présents dans les données.

		même taille de cluster							tailles de cluster différentes						
		2	3	4	5	6	7	8	2	3	4	5	6	7	8
$p = 0.9$	2	**100**	0	0	0	0	0	0	**100**	0	0	0	0	0	0
	3	0	**100**	0	0	0	0		0	**100**	0	0	0	0	0
	4	0	0	**99**	0	1	0	0	0	6	**85**	5	3	1	0
	5	0	0	2	**98**	0	0	0	0	3	34	**50**	8	4	1
	6	0	0	0	8	**85**	6	1	0	0	29	49	**15**	6	1
	7	0	0	0	1	24	**56**	19	0	0	30	50	13	**6**	1
$p = 0.6$	2	**100**	0	0	0	0	0	0	**100**	0	0	0	0	0	0
	3	0	**100**	0	0	0	0	0	0	**99**	1	0	0	0	0
	4	0	0	**99**	1	0	0	0	0	14	**68**	9	7	2	0
	5	0	0	4	**79**	14	1	2	0	18	50	**22**	4	6	0
	6	0	0	1	22	**49**	22	6	0	20	46	16	**13**	4	1
	7	0	0	0	16	47	**24**	13	0	22	56	14	5	**3**	0
$p = 0.5$	2	**100**	0	0	0	0	0	0	**98**	2	0	0	0	0	0
	3	0	**98**	2	0	0	0	0	1	**91**	7	0	1	0	0
	4	0	0	**87**	9	3	1	0	1	43	**32**	16	4	1	3
	5	0	0	15	**44**	26	12	3	2	34	44	**9**	8	3	0
	6	0	1	11	28	**22**	25	13	0	47	32	15	**5**	1	0
	7	0	0	6	34	28	**17**	15	2	30	46	14	5	**3**	0

Table 2.2 Matrice de confusion. Nombre estimé de clusters (colonnes) $Q_{ILosbm} \in \{2, \ldots, 8\}$ versus vrai nombre de clusters (lignes) $Q_{Vrai} \in \{2, \ldots, 7\}$ calculé pour des réseaux à même taille de cluster ou tailles de cluster différentes, et trois probabilités intra cluster $p \in \{0.9, 0.6, 0.5\}$.

2.4 Caractérisation des sous-graphes

Dans de nombreux cas, les méthodes basées sur des extensions du modèle SBM donnent des résultats déjà connus du praticien. À titre illustratif, nous avons caractérisé récemment (Zreik, Latouche et Bouveyron, 2015b) une grande partie des échanges maritimes dans le monde à l'aide d'un réseau. Les clusters obtenus sur ce réseau à l'aide, en autre, de l'algorithme VBEM pour SBM décrit en section 2.2, donnent exactement les aires géographiques bien connues des géographes. Nous avons donc cherché dans Jernite, Latouche et al., 2014 à construire un nouveau modèle permettant d'intégrer une information connue par le praticien sous la forme une partition des nœuds donnée. L'objectif est alors de chercher des clusters de nœuds dont la topologie est non triviale. Ce travail était également motivé par une collaboration avec des historiens du LAMOP ayant compilé, sous la forme d'un réseau, certaines interactions sociales dans la Gaulle Mérovingienne du 6ème siècle. Ce réseau est

constitué d'arcs catégoriels, et non plus simplement binaires.

2.4.1 Modèle

Nous supposons ici que nous disposons d'une partition connue des nœuds. Notons que cette partition induit une décomposition du réseau en sous-graphes. Un sous-graphe est ainsi constitué des nœuds du groupe associé et de tous les arcs entre ces nœuds du réseau. Le modèle est ici présenté dans un cas orienté.

Le tirage aléatoire du graphe se fait en trois étapes. Tout d'abord, un arc binaire est généré pour chaque paire de nœud (i, j) en fonction des sous-graphes s_i et s_j auxquels ils appartiennent :

$$A_{ij}|\gamma_{s_i,s_j} \sim \mathcal{B}(\gamma_{s_i,s_j}).$$

Cette étape est très particulière et est motivée exclusivement par des considérations techniques liées au réseau mérovingien à analyser. Dans l'application publiée dans Jernite, Latouche et al., 2014, les sous-graphes correspondent en effet à différentes provinces. Pour des raisons évidentes de transport, il est donc attendu que, globalement, les nœuds de provinces proches interagissent plus que des nœuds de provinces éloignées géographiquement. Ensuite, chaque nœud est associé à un vecteur latent Z_i à partir d'une loi multinomiale dont le paramètre dépend du sous-graphe connu de i :

$$Z_i|\pi_{s_i} \sim \mathcal{M}(1, \pi_{s_i}),$$

avec $\sum_{q=1}^{Q} \alpha_{sq} = 1$ pour tous les sous-graphes s. Tous les Z_i sont tirés de manière indépendante. Enfin, les clusters et la présence des arcs étant déterminés, le type (parmi C) de chaque arc est généré aléatoirement suivant une loi multinomiale conditionnelle :

$$X_{ij}|Z_{iq}Z_{jr}A_{ij} = 1, \mu \sim \mathcal{M}(1, \mu_{qr}),$$

où $\sum_{c=1}^{C} \mu_{qrc} = 1, \forall (q, r) \in 1, .., Q^2$. Ainsi, X_{ij} est un vecteur binaire et $X_{ijc} = 1$ indique que l'arc (i, j) est de type c. Nous avons appelé le modèle ainsi constitué modèle à sous-graphes aléatoires, ou random subraph model (RSM) en anglais.

Notons que la distinction faite ici entre la génération de la présence des arcs à partir des sous-graphes uniquement, et du type à partir des clusters, eux mêmes dépendant des sous-graphes, est abandonnée dans la version dynamique de RSM présentée dans le chapitre suivant.

2.4.2 Inférence

RSM est un modèle dérivé du modèle SBM et souffre par conséquence de contraintes héritées. En particulier, le calcul de la vraisemblance des données a un coût exponentiel ce qui ne permet pas d'utiliser les critères classiques de sélection de modèles pour estimer le nombre de clusters. Nous avons donc considéré un cadre Bayésien pour RSM à l'image de nos travaux pour SBM et OSBM. De plus, la loi des états cachés, sachant les données et les paramètres, n'a pas de forme analytique.

Des lois a priori conjuguées ont été utilisées pour caractériser l'a priori sur les paramètres. Ainsi,

$$p(\gamma_{qr}) = \text{Beta}(\gamma_{qr};\, a^0_{qr}, b^0_{qr}), \forall (q, r) \in \{1, \dots, S\}^2,$$
$$p(\alpha_s) = \text{Dir}\left(\alpha_s;\, \chi^0_{s1}, \dots, \chi^0_{sQ}\right), \forall s \in \{1, \dots, S\},$$

et

$$p(\mu_{qr}) = \text{Dir}\left(\mu_{qr};\, \Xi^0_{qr1}, \ldots, \Xi^0_{qrC}\right), \forall (q,r) \in \{1, \ldots, Q\}^2.$$

Pour une discussion sur le choix des hyperparamètres, nous renvoyons le lecteur à la section 2.2.

Le modèle RSM ne faisant par intervenir de fonction de lien non linéaire, un algorithme VBEM classique peut directement être utilisé afin de maximiser une borne inférieure de la log-vraisemblance marginale $\log p(A, X|Q)$ des données, par rapport à une loi r, approximation de la loi a posteriori $p(Z, (\pi_s)_s, \gamma, \mu|A, X, Q)$. À convergence, r permet notamment d'obtenir le clustering des nœuds alors que la borne de l'algorithme VBEM est utilisée comme approximation variationnelle de la log-vraisemblance marginale. Un ensemble de valeurs de Q est considéré et la valeur de Q maximisant la borne inférieure est retenue comme estimation du nombre de clusters présents dans les données. Pour l'optimisation fonctionnelle, nous avons considéré la famille de loi factorisable suivante :

$$r(Z, (\pi_s)_s, \gamma, \mu) = \left(\prod_{i=1}^{n} r(Z_i)\right)\left(\prod_{s=1}^{S} r(\alpha_s) \prod_{r=1}^{S} r(\gamma_{sr})\right) \prod_{q,r}^{Q} r(\mu_{qr}).$$

L'optimisation étant standard, nous ne présentons ici que le critère de sélection de modèles, c'est-à-dire la borne à convergence de l'algorithme VBEM.

Proposition 2.7. *À convergence de l'algorithme VBEM, la borne inférieure variationnelle de la log-vraisemblance marginale du modèle RSM est donnée par :*

$$\mathcal{L}_Q(r) = \sum_{q,r}^{S} \log\left(\frac{B(a_{qr}, b_{qr})}{B(a^0_{qr}, b^0_{qr})}\right) + \sum_{s=1}^{S} \log\left(\frac{C(\chi_s)}{C(\chi^0_s)}\right) + \sum_{qr}^{Q} \log\left(\frac{C(\Xi_{qr})}{C(\Xi^0_{qr})}\right) - \sum_{i=1}^{n}\sum_{q=1}^{Q} \tau_{iq} \log(\tau_{iq}),$$

où $C(x) = \frac{\prod_{d=1}^{D} \Gamma(x_d)}{\Gamma(\sum_{d=1}^{D} x_d)}$ *si* $x \in \mathbb{R}^D$, $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \forall (a,b) \in \mathbb{R}^2$, *et* $\Gamma(\cdot)$ *est la fonction gamma. Les paramètres* a_{qr}, b_{qr}, χ_s, Ξ_{qr} *correspondent aux hyperparamètres* a^0_{qr}, b^0_{qr}, χ^0_s, Ξ^0_{qr} *mis à jour en fonction des données, au cours de l'optimisation.* τ_{iq} *est l'approximation de la probabilité que le nœud* i *appartienne au cluster* q. *Ces termes sont donnés dans l'appendice de* Jernite, Latouche *et al., 2014.*

2.4.3 Expérimentations numériques

Nous présentons dans cette section une des expériences que nous avons réalisée dans Jernite, Latouche et al., 2014 sur données simulées afin de tester la méthodologie. 50 réseaux à $n = 100$ nœuds ont ainsi été générés selon un modèle RSM à trois clusters et trois types d'arêtes. Les nœuds d'un même cluster se connectent principalement avec des arcs de type 1 contrairement aux nœuds de clusters différents qui se connectent principalement avec des arcs de type 3. La procédure d'écrite précédemment est utilisée sur chaque réseau afin d'estimer le nombre de clusters et de proposer un clustering des nœuds. La pertinence de l'estimation est évaluée à l'aide du critère ARI. Le critère évalue donc ici à la fois l'estimation de Q et de Z. Nous observons sur la figure 2.4 que le critère de sélection de modèles atteint bien son maximum pour $Q = 3$ clusters. Les plus grandes valeurs d'ARI sont obtenues également pour $Q = 3$ avec une médiane proche de 1.

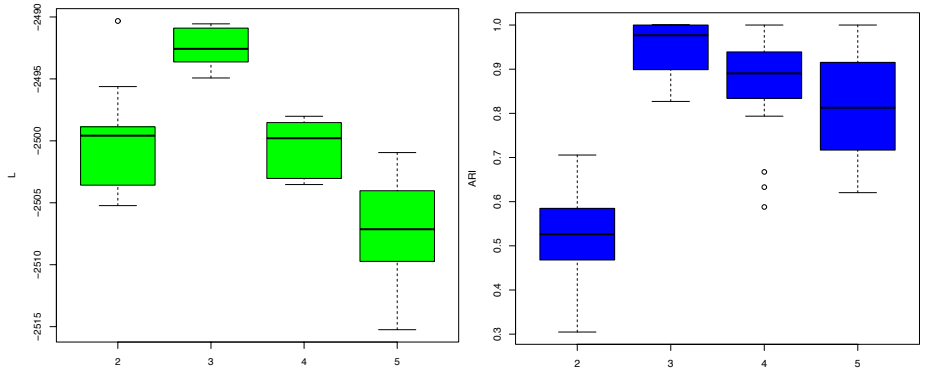


Figure 2.4 – Figure de gauche : répartition des valeurs du critère de sélection de modèles proposé en fonction du nombre de clusters. Figure de droite : répartition des valeurs d'ARI. 50 réseaux analysés à $Q = 3$ clusters.

2.5 Analyse conjointe de réseaux et de textes

Les méthodes de clustering de nœuds, qu'elles soient basées sur un modèle probabiliste ou pas, sont fortement contraintes quant au type de réseau pouvant être analysé. En effet, l'immense majorité n'autorise que l'étude de réseaux binaires où seule la présence ou l'absence de connexions entre les nœuds est retenue comme source de données. De nombreuses extensions ont été proposées ces dix dernières années afin d'autoriser d'autres types d'arêtes ou d'arcs. Nous pouvons par exemple citer les travaux de Mariadassou, Robin et Vacher, 2010 modélisant les connexions discrètes à l'aide de lois de Poisson conditionnelles. Comme nous l'avons vu, le modèle RSM décrit dans la section précédente autorise lui les connexions catégorielles.

En sciences sociales, il est courant d'observer des échanges entre des individus, ces échanges pouvant être représentés sous la forme d'un réseau. Que ce soit à travers l'écriture d'emails, de messages, ou de posts, ces interactions font la plupart du temps intervenir du texte. En pratique, l'envoi d'un ou plusieurs emails d'un individu i à un individu j est souvent enregistré uniquement sous la forme d'un arc binaire indiquant la présence d'une connexion entre i et j. Il est également fréquent de considérer le nombre d'emails envoyés, l'arc étant ainsi associé à un poids discret. Dans le chapitre 3, nous verrons comment modéliser la dynamique de l'évolution de ces poids. Dans cette section, nous présentons nos travaux publiés dans Bouveyron, Latouche et Zreik, 2016 permettant d'associer à l'arc (i, j) tous les emails, ou plus généralement l'ensemble des documents, envoyés de i à j.

Le modèle probabiliste, appelé stochastic topic block model (STBM), fait intervenir deux étapes principales. La première a pour objectif la fabrication des connexions. La seconde vise à construire des documents pour chaque connexion présente. Le modèle est ici présenté dans le cas de réseaux orientés.

2.5.1 Modélisation de la présence de connexions

Un modèle SBM standard est considéré pour générer la présence ou l'absence de connexions entre les nœuds d'un graphe. Ainsi, chaque nœud i est associé à un vecteur binaire Y_i tiré à partir d'une loi multinomiale de paramètres $(1, \pi)$, où π caractérise le poids des clusters. Ce vecteur est noté Y_i ici et non pas Z_i, contrairement aux autres sections, afin de pouvoir clairement faire la distinction entre les clusters de nœuds et les clusters de mots. Ensuite,

pour chaque paire de nœuds (i, j), un arc X_{ij} est généré aléatoirement selon une loi de Bernoulli de paramètre μ_{qr} si i appartient au cluster q et j à r. Pour plus de détails, nous renvoyons le lecteur à la section 2.2 de ce chapitre.

2.5.2 Modélisation des documents

Afin de modéliser la construction des documents associés aux arcs présents, nous nous sommes appuyés sur le modèle d'allocation latente de Dirichlet, latent Dirichlet allocation (LDA) (Blei, Ng et Jordan, 2003) en anglais. Ce dernier fait référence en analyse statistique de textes. Les méthodes d'inférence associées permettent en particulier d'identifier les thèmes principaux des documents d'un corpus.

Dans notre cadre, pour chaque paire (q, r) de clusters de nœuds, un vecteur θ_{qr} est ainsi tiré aléatoirement selon une loi de Dirichlet :

$$\theta_{qr} \sim \text{Dir}\left(\alpha = (\alpha_1, \ldots, \alpha_K)\right),$$

donc $\sum_{k=1}^{K} \theta_{qrk} = 1$. Ces tirages sont tous *iid*. Dans la littérature, il existe de nombreuses discussions au sujet du vecteur α. Lors de l'inférence, il est en effet possible de construire une estimation de α à l'aide d'une procédure de type Bayes empirique. Malheureusement, dans le cadre de LDA, les méthodes d'optimisation existantes donnent des résultats très instables, en particulier vis à vis de l'initialisation. Pour espérer obtenir une estimation pertinente de α, il faut donc répéter toute l'optimisation un très grand nombre de fois. Pour ces raisons, et parce qu'il n'existe pas à notre connaissance de travaux théoriques complets concernant le rôle de ce vecteur dans LDA, nous avons choisi de fixer toutes ses composantes à 1. La loi de Dirichlet correspond alors à la loi uniforme sur le simplexe. La probabilité θ_{qrk} encode le poids du thème de discussion k parmi K dans les échanges entre les nœuds du cluster q vers ceux dans r.

Par la suite, nous notons $W_{ij} = (W_{ij}^d)_d$ l'ensemble des documents envoyés de i à j et $\theta = (\theta_{qr})_{qr}$. Le $n-$ième mot W_{ij}^{dn} du document d dans W_{ij} est alors associé à un vecteur binaire Z_{ij}^{dn} supposé tiré selon une loi multinomiale conditionnelle :

$$Z_{ij}^{dn} | Y_{iq} Y_{jr} X_{ij} = 1, \theta \sim \mathcal{M}(1, \theta_{qr}). \tag{2.1}$$

Dès lors, si un arc est présent entre les nœuds (i, j), et que le nœud i appartient au cluster q et j à r, le mot W_{ij}^{dn} appartient au cluster de mots k avec probabilité θ_{qrk}. Un cluster de mots est vu ici comme un thème de discussion. Ensuite, conditionnellement à Z_{ij}^{dn}, le mot W_{ij}^{dn} lui même est généré à partir d'une loi multinomiale :

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1, \beta \sim \mathcal{M}\left(1, \beta_k\right), \tag{2.2}$$

où β_k est un vecteur de probabilités à V composantes, telles que $\sum_{v=1}^{V} \beta_{kv} = 1$, où V désigne la taille du dictionnaire utilisé. Tous les vecteurs β_k sont rangés dans une matrice notée $\beta = (\beta_k)_k$. Ainsi, $\sum_{v=1}^{V} W_{ij}^{dnv} = 1$ et $W_{ij}^{dnv} = 1$ signifie que le mot v du dictionnaire apparait en position n du document d.

La description de cette partie du modèle probabiliste est fidèle à celle de Bouveyron, Latouche et Zreik, 2016, elle même cohérente avec la présentation de LDA dans Blei, Ng et Jordan, 2003. Nous proposons maintenant dans cette section de réécrire le procédé de tirage associé à (2.1) et (2.2) afin de mieux illustrer le modèle sous jacent. Une marginalisation de (2.2) à partir de (2.1) donne :

$$W_{ij}^{dn} | Y_{iq} Y_{jr} X_{ij} = 1, \theta, \beta \sim \sum_{k=1}^{K} \theta_{qrk} \mathcal{M}\left(1, \beta_k\right). \tag{2.3}$$

Tous les mots W_{ij}^{dn} de W_{ij} sont tirés de manière iid et selon (2.3), le mot v du dictionnaire apparait à chaque fois avec probabilité :

$$\begin{aligned}\mathbb{P}(W_{ij}^{dnv} = 1 | Y_{iq}Y_{jr}X_{ij} = 1, \theta, \beta, K, Q) &= \sum_{k=1}^{K} \theta_{qrk}\beta_{kv} \\ &= \theta_{qr}^{\intercal}\beta_{\cdot v},\end{aligned}$$

où $\beta_{\cdot v}$ désigne la colonne v de la matrice β. Ainsi, d'un point de vue modèle, seul compte le nombre de fois où un mot du dictionnaire est présent, pas sa position. Notre modélisation s'inscrit donc dans la famille de modèles de type sacs de mots, ou bags of words en anglais. En notant $\tilde{W}_{ij}$ le vecteur de taille V où $\tilde{W}_{ij}^{v}$ indique le nombre de fois où v est présent dans W_{ij}, le modèle dans Bouveyron, Latouche et Zreik, 2016 donne :

$$\tilde{W}_{ij} | Y_{iq}Y_{jr}X_{ij} = 1, \theta, \beta \sim \mathcal{M}(N_{ij}, \theta_{qr}\beta), \tag{2.4}$$

où N_{ij} désigne le nombre total de mots échangés entre les nœuds i et j. Comme dans le cadre de LDA, chaque N_{ij} est supposé connu et fixe. Il est également possible d'ajouter une nouvelle couche au modèle probabiliste en tirant ces comptages à l'aide de loi de Poisson par exemple. La caractérisation (2.4) a pour avantage de mettre en avant les données qui seront finalement à analyser. Ainsi, pour chaque paire de nœuds (i, j), la présence d'un arc est stockée dans X_{ij} et si $X_{ij} = 1$ alors les comptages des mots échangés sont également enregistrés dans $\tilde{W}_{ij}$.

2.5.3 Pivot

Ci-dessous, $Y = (Y_i)_i$, $Z = (Z_{ij}^{dn})_{ijdn}$, et $W = (\tilde{W}_{ij})_{ij}$. La loi jointe de STBM se factorise de la manière suivante :

$$p(X, W, Y, Z, \theta | \pi, \mu, \beta, K, Q) = p(W, Z, \theta | X, Y, \beta, K, Q) p(X, Y | \pi, \mu, Q). \tag{2.5}$$

Le terme de droite $p(X, Y|\pi, \mu, Q)$ dans (2.5) correspond exactement à la vraisemblance des données complétées pour le modèle SBM. De plus, à Y fixé, tous les documents $\tilde{W}_{ij}$ pour lesquels i est dans le cluster q et j dans r, sont indépendants et suivent la même loi (2.4). Il est alors possible de construire des agrégations de données en comptant combien de fois un mot v du dictionnaire est apparu en tout, pour toutes les paires (i, j) telles que $Y_{iq}Y_{jr}X_{ij} = 1$. En notant $\tilde{\tilde{W}}_{qr}$ le vecteur de taille V des comptages associés, nous trouvons :

$$\tilde{\tilde{W}}_{qr} | \theta, \beta \sim \mathcal{M}(\tilde{N}_{qr}, \theta_{qr}\beta), \tag{2.6}$$

où $\tilde{N}_{qr} = \sum_{i,j,Y_{iq}Y_{jr}X_{ij}=1} N_{ij}$. Les Q^2 documents $\tilde{\tilde{W}}_{qr}$ ainsi construits suivent alors exactement le modèle LDA standard de loi jointe $p(W, Z, \theta | X, Y, \beta, K, Q)$. La matrice Y est donc pivot puisqu'elle fait le lien entre une partie modélisation de graphes et une partie modélisation de textes.

2.5.4 Inférence

La stratégie d'inférence que nous avons développée s'appuie largement sur la propriété de pivot de Y discutée à la section précédente. Ainsi, nous considérons le problème d'optimisation de $\log p(X, W, Y | \pi, \mu, \beta, K, Q)$. Ce terme ne peut pas être évalué de manière explicite

de par la structure du modèle graphique. Pour plus de détails, nous renvoyons le lecteur à la section 2.2 de ce chapitre. En revanche, une décomposition variationnelle simple donne :

$$\log p(X, W, Y|\pi, \mu, \beta, K, Q) = \mathcal{L}_{K,Q}\left(r; Y, \pi, \mu, \beta\right) + \mathrm{KL}\left(r \parallel p(\cdot|X, W, Y, \pi, \mu, \beta, K, Q)\right),$$

où la borne inférieur peut s'écrire :

$$\mathcal{L}_{K,Q}(r; Y, \pi, \mu, \beta) = \tilde{\mathcal{L}}_{K,Q}\left(r; Y, \beta\right) + \log p(X, Y|\pi, \mu, Q).$$

À Y fixe et une fois les documents agrégés, $\tilde{\mathcal{L}}_{K,Q}$ correspond exactement à la borne de l'algorithme VEM pour LDA introduit par Blei, Ng et Jordan, 2003. Cet algorithme peut donc être utilisé directement afin d'estimer r et β. Les paramètres π et μ s'obtiennent directement à partir de Y. Seule l'optimisation en fonction de Y pose réellement difficulté. En effet, chercher l'optimum global en testant toutes les Q^n configurations de Y n'est pas faisable. C'est un problème combinatoire. Nous proposons ici une heuristique simple à travers un algorithme glouton. Les nœuds sont vus chacun à leur tour. Si un ou plusieurs changements de clusters de i améliorent la borne inférieure, alors celui induisant le plus grand accroissement est appliqué. Sinon, le cluster du nœud reste inchangé. Toutes ces étapes d'optimisation sont répétées jusqu'à convergence de la borne variationnelle.

2.5.5 Sélection de modèles

A l'aide de deux approximations de Laplace, d'une estimation variationnelle, et de la formule de Stirling, nous avons dérivé un critère ICL pour le modèle STBM.

Proposition 2.8. *La vraisemblance jointe* $\log p(X, W, Y|K, Q)$ *du modèle STBM peut être approchée par un critère de type ICL de la forme :*

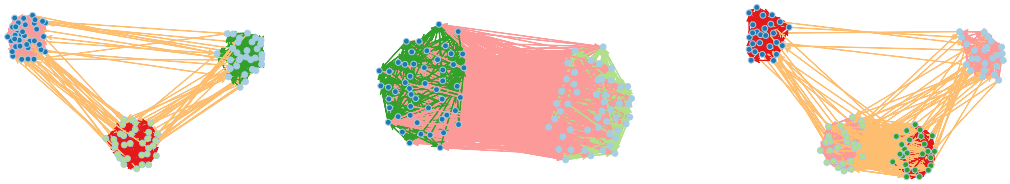
$$ICL_{STBM} = \tilde{\mathcal{L}}_{K,Q}(r; Y, \beta) - \frac{K(V-1)}{2}\log Q^2 + \max_{\pi,\mu} \log p(X, Y|\pi, \mu, Q) - \frac{Q^2}{2}\log n(n-1) - \frac{Q-1}{2}\log n.$$

2.5.6 Expérimentations numériques

Plusieurs études ont été réalisées dans Bouveyron, Latouche et Zreik, 2016 afin de tester sur données simulées la procédure d'inférence. Nous présentons ici les résultats concernant l'évaluation de la capacité de la méthode à retrouver les clusters de nœuds et les thèmes de discussion, à K et Q fixés.

Trois scénarios ont d'abord été fixés pour la génération de réseaux selon le modèle STBM. Les paramètres sont donnés dans le tableau 2.3 et des exemples sont proposés dans la figure 2.5. À partir des ces trois scénarios, nous avons considéré trois situations : situation de référence correspondant exactement aux paramètres indiqués dans le tableau 2.3 ("Facile"), une situation ("Dure 1") où les clusters de nœuds ont plus de connexions vers les autres clusters ($\pi_{q\neq r} = 0.2$ sauf pour B), et une situation ("Dure 2") où 40% des messages sont bruités, *i.e.* générés à partir d'un autre thème de discussion. Pour chaque scénario et chaque situation, 20 réseaux ont été simulés et analysés.

SBM et LDA ont été considérés, comme référence, pour l'analyse (binaire) des réseaux et des textes respectivement. Les résultats sont présentés dans le tableau 2.4. La méthode d'inférence associée à STBM est la seule à pouvoir analyser à la fois la présence d'arêtes dans un réseau et les textes associés. Les ARI obtenus illustrent la pertinence de l'approche.



Scénario A Scénario B Scénario C

Figure 2.5 – Exemples de réseau simulés selon les scénarios A, B, et C.

Scénario	A	B	C
n	100		
K	4	3	3
Q	3	2	4
π	$(1/Q, ..., 1/Q)$		
μ	$\begin{cases} \mu_{qq} = & 0.25 \\ \mu_{qr,\, r \neq q} = & 0.01 \end{cases}$	$\mu_{qr, \forall q,r} = 0.25$	$\begin{cases} \mu_{qq} = & 0.25 \\ \mu_{qr,\, r \neq q} = & 0.01 \end{cases}$
θ	$\begin{cases} \theta_{111} = \theta_{222} = & 1 \\ \theta_{333} = & 1 \\ \theta_{qr4,\, r \neq q} = & 1 \\ \text{sinon} & 0 \end{cases}$	$\begin{cases} \theta_{111} = \theta_{222} = & 1 \\ \theta_{qr3,\, r \neq q} = & 1 \\ \text{sinon} & 0 \end{cases}$	$\begin{cases} \theta_{111} = \theta_{331} = & 1 \\ \theta_{222} = \theta_{442} = & 1 \\ \theta_{qr3,\, r \neq q} = & 1 \\ \text{sinon} & 0 \end{cases}$

Table 2.3 Valeurs des paramètres pour les trois scénarios considérés.

		Scénario A		Scénario B		Scénario C	
Facile	Méthode	nœud ARI	thème ARI	nœud ARI	thème ARI	nœud ARI	thème ARI
	SBM	1.00±0.00	–	0.01±0.01	–	0.69±0.07	–
	LDA	–	0.97±0.06	–	1.00±0.00	–	1.00±0.00
	STBM	0.98±0.04	0.98±0.04	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00

		Scénario A		Scénario B		Scénario C	
Dure 1	Méthode	nœud ARI	thème ARI	nœud ARI	thème ARI	nœud ARI	thème ARI
	SBM	0.01±0.01	–	0.01±0.01	–	0.01±0.01	–
	LDA	–	0.90±0.17	–	1.00±0.00	–	0.99±0.01
	STBM	1.00±0.00	0.90±0.13	1.00±0.00	1.00±0.00	1.00±0.00	0.98±0.03

		Scénario A		Scénario B		Scénario C	
Dure 2	Méthode	nœud ARI	thème ARI	nœud ARI	thème ARI	nœud ARI	thème ARI
	SBM	1.00±0.00	–	-0.01±0.01	–	0.65±0.05	–
	LDA	–	0.21±0.13	–	0.08±0.06	–	0.09±0.05
	STBM	0.99±0.02	0.99±0.01	0.59±0.35	0.54±0.40	0.68±0.07	0.62±0.14

Table 2.4 Résultats de clustering pour SBM, LDA, et STBM. 20 réseaux analysés pour trois scénarios A, B, C, et trois situations "Facile", "Dure 1", "Dure 2". Les valeurs moyennes d'ARI sont reportées ainsi que les écarts types, pour le clustering des nœuds et des textes.

Conclusion

Dans ce chapitre, nous avons proposé deux méthodes permettant de réaliser l'inférence du modèle SBM. La seconde permet notamment d'estimer le nombre de clusters présents dans les données ainsi que la partition des nœuds associée. Le coût algorithmique de l'algorithme correspondant est également sensiblement plus faible que celui des autres techniques basées sur SBM. Rappelons que ce modèle fait référence dans la littérature statistique en analyse des réseaux. Pourtant, bien que capable de caractériser des structures de natures différentes, comme des communautés, des schémas en étoile, et des hubs, il est limité sur plusieurs aspects. Tout d'abord, chaque nœud ne peut appartenir qu'à un seul cluster. Nous avons donc proposé le modèle OSBM autorisant la recherche de clusters chevauchants ainsi que deux procédures d'inférence. Le cadre Bayésien de la deuxième fait apparaître naturellement un critère de sélection de modèles. Ensuite, des études sur données réelles ont montré que les clusters identifiés par des méthodes basées sur SBM pouvaient correspondre à des structures déjà connues des praticiens. Le modèle RSM a donc été proposé afin d'intégrer dans la modélisation une pré-classification des nœuds fournie. L'inférence permet alors d'identifier des clusters porteurs d'une information non expliquée par la pré-classification. Enfin, nos travaux ont donné lieu au modèle STBM. L'inférence de ce modèle autorise l'analyse conjointe d'un réseau et d'un ensemble de documents portés par les connexions.

3

Des réseaux, des textes, des blocs. Modèles dynamiques

3.1 Introduction . . . 34
3.2 Temps discret . . . 34
3.2.1 Evolution des sous-graphes 35
3.2.1.a Modèle 35
3.2.1.b Inférence 36
3.2.1.c Expérimentations numériques 37
3.2.2 Construction d'une partition du temps 38
3.2.2.a Inférence 39
3.2.2.b Expérimentations numériques 40
3.3 Temps continu et segmentation . . . 41
3.3.1 Processus ponctuels et comptages 42
3.3.2 Points de rupture 43
3.3.3 Vraisemblance pénalisée et approximations variationnelles 43
3.3.4 Optimisation 44
3.3.5 Inférence 45
3.3.6 Expérimentations numériques 46

Cette seconde thématique de recherche s'intéresse à l'analyse de données d'interactions. Une caractéristique clé de ces données est leur caractère dynamique. Un temps discret peut alors être considéré pour décrire les changements temporels. D'autres processus permettent également de traiter les données à travers un temps continu. Ces deux sous-axes sont discutés ci-dessous. Le premier a donné lieu à quatre articles dans des journaux (Zreik, Latouche et Bouveyron, 2015a ; Zreik, Latouche et Bouveyron, 2016 ; Corneli, Latouche et Rossi, 2016a ; Corneli, Latouche et Rossi, 2016b). Mes travaux dans le second se sont traduits par un article de journal (Corneli, Latouche et Rossi, à paraître).

3.1 Introduction

Les développements des moyens de communication et de l'aire numérique plus généralement favorisent le stockage de plus en plus important de données de type interaction. Les interactions peuvent par exemple correspondre à un ensemble d'échanges d'emails entre des employés d'une entreprise. Ils peuvent également être caractérisés par des posts sur des réseaux sociaux tels que Twitter, Facebook, ou Linkedin. En pratique, une interaction se fait à un instant temporel précis stocké dans une base de données. Si un individu i interagit avec j, au temps ν, alors l'interaction correspondante peut être représentée sous la forme d'un triplet (i, j, ν). L'ensemble de tous les triplets peut alors être utilisé pour construire un réseau dynamique où chaque individu est associé à un nœud, et un arc ou arête entre deux nœuds est présent au temps ν si l'interaction (i, j, ν) est enregistrée dans les données.

De nombreux modèles probabilistes ont été proposés pour l'étude des réseaux dynamiques ces 10 dernières années. La plupart sont des modèles à temps discret. Autrement dit, des intervalles de temps prédéfinis sont considérés et les interactions ayant lieu durant ces intervalles sont agrégées afin d'obtenir des images des réseaux, à temps fixes. Dans le cas binaire, deux nœuds sont connectés dans l'image d'un réseau, si au moins une interaction a lieu dans l'intervalle de temps associé. Comme nous le verrons par la suite, il est également possible d'étudier le nombre ou le type d'interactions ayant lieu dans chaque intervalle. Les modèles les plus prometteurs travaillent en réalité à temps continu et font intervenir des processus ponctuels. Les modèles à temps discret sont en effet contraints. Comme discuté en particulier dans Matias, Rebafka et Villers, 2015, le choix des intervalles de temps, pour la construction des images, peut avoir un impact fort sur les résultats d'analyse. Nous reviendrons sur cette question à la fin de ce chapitre.

Ci-dessous, nous désignons par $\mathcal{E} = \{(i_m, j_m, \nu_m)\}_{1 \leq m \leq M}$ l'ensemble des M interactions observées. $\mathcal{E}$ est un sous-ensemble de $\{1, \ldots, n\}^2 \times \mathbb{R}^+$ et la période d'étude correspond à l'intervalle de temps $[0, T]$. Sans perte de généralité, $\mathcal{E}$ est supposé ordonné par la variable de temps. De plus, chaque temps d'interaction est unique. Dès lors $0 \leq \nu_1 < \nu_2 < \ldots < \nu_M \leq T$. L'ensemble des temps d'interaction ayant lieu uniquement entre les nœuds i et j est ensuite noté :

$$\mathcal{A}^{(i,j)} := \{\nu_1^{(i,j)}, \ldots, \nu_{M^{(i,j)}}^{(i,j)}\}, \tag{3.1}$$

où $\nu_1^{(i,j)} < \nu_2^{(i,j)} < \cdots < \nu_{M^{(i,j)}}^{(i,j)}$. À temps discret, une décomposition de $[0, T]$ en U sous-intervalles $I_u =]t_{u-1}, t_u]$ est considérée, délimités par les temps :

$$0 = t_0 < t_1 < \cdots < t_U = T.$$

Finalement, nous notons $X = (X^{(u)})_u$ l'ensemble des matrices d'adjacence codant les images du réseau pour tous les intervalles I_u.

3.2 Temps discret

En temps discret, nous nous sommes intéressés à deux types de problèmes. Le premier vise à modéliser un réseau dynamique lorsqu'une partition des nœuds en sous-graphes est fournie. Ces travaux constituent une extension du modèle RSM présenté en section 2.4. Le second a pour objectif d'identifier des clusters d'intervalles temporels.

3.2.1 Evolution des sous-graphes

Le modèle RSM introduit en section 2.4 permet de chercher des clusters lorsque le praticien dispose dès le départ d'une partition des nœuds. Cette dernière induit une décomposition du réseau en sous-graphes où chaque sous-graphe est constitué des nœuds du groupe connu et de tous les arcs entre ces nœuds du réseau. La partition donnée étant intégrée par le modèle, l'inférence associée autorise en pratique la recherche de clusters non triviaux. Nous présentons ici notre extension (Zreik, Latouche et Bouveyron, 2016) du modèle RSM au cadre dynamique. La méthodologie est introduite pour des réseaux orientés.

3.2.1.a Modèle

Pour des considérations géographiques liées au réseau historique ayant motivé le développement de RSM, ce dernier fait la distinction entre la création d'arcs et la génération du type associé. Pour la version dynamique, appelée dynamic RSM en anglais, nous avons relâché cette contrainte. Ainsi, au temps t, l'absence d'arc entre les nœuds (i, j) et le type de l'arc si présent sont tous les deux caractérisés par le vecteur binaire $X_{ij}^{(t)}$. $X_{ij}^{(t)}$ code pour 0 en cas d'absence au temps t. Sinon, le vecteur code pour un type parmi C. Une loi multinomiale conditionnelle est utilisée pour générer $X_{ij}^{(t)}$:

$$X_{ij}^{(t)} | Z_{iq}^{(t)} Z_{jr}^{(t)} = 1, \mu \sim \mathcal{M}(1, \mu_{qr}),$$

où $\mu = (\mu_{qr})_{qr}$ et $\sum_{c=0}^{C} \mu_{qrc} = 1, \forall q, r$. Contrairement à toutes les approches décrites dans le chapitre précédant, le vecteur $Z_i^{(t)}$ d'appartenance aux clusters est maintenant indicé par le temps. Ainsi, pour chaque temps t, $\sum_{q=1}^{Q} Z_{iq}^{(t)} = 1$ et $Z_{iq}^{(t)} = 1$ indique que le nœud i appartient au cluster q au temps t. Chaque nœud peut donc changer de clusters au cours du temps. À l'inverse, les paramètres de connectivité dans μ sont supposés fixes. Ces choix sont à mettre en perspective par rapport aux travaux récents de Matias et Miele, à paraître. Ils ont montré que pour respecter des contraintes d'identifiabilité dans des modèles dérivés de SBM, le contenu des clusters ou les paramètres de connectivité ne peuvent pas ensemble être fonctions du temps.

Chaque vecteur $Z_{iq}^{(t)}$ est supposé issu d'une loi multinomiale :

$$Z_i^{(t)} | \pi_{s_i}^{(t)} \sim \mathcal{M}(1, \pi_{s_i}^{(t)}),$$

où $s(i)$ désigne le sous-graphe connu de i, comme en section 2.4. Le sous-graphe s est donc modélisé par des vecteurs $\pi_s^{(t)}$ de proportions des clusters et $\sum_{q=1}^{Q} \pi_{sq}^{(t)} = 1$ pour chaque t. Ainsi, la dynamique du réseau est d'abord caractérisée par l'évolution du poids des clusters dans les sous-graphes. Les clusters sont associés à des propriétés de connexion, et c'est leur présence au sein des sous-graphes qui explique la construction des arcs, au cours du temps. Afin de capturer la dynamique temporelle des vecteurs $\pi_s^{(t)}$, un modèle linéaire dynamique, linear dynamical model (LDM) en anglais, a été utilisé. Les termes $\pi_{sq}^{(t)}$ étant des probabilités, LDM est appliqué sur une fonction non linéaire de ces derniers :

$$\pi_s^{(t)} = f(\gamma_s^{(t)}),$$

où $f : \mathbb{R}^Q \to [0, 1]^Q$ telle que :

$$\pi_{sq}^{(t)} = \exp(\gamma_{sq}^{(t)} - C(\gamma_s^{(t)})), \forall s, q, t,$$

et $\gamma_{sQ}^{(t)} = 0$. Ici $C(\cdot)$ est un terme de normalisation assurant que $\sum_{q=1}^{Q} \pi_{sq}^{(t)} = 1, \forall s, t$. Le choix arbitraire de fixer la dernière composante du vecteur $\gamma_s^{(t)}$ à 0 est largement retenu dans la littérature. Il permet de s'assurer que $f(\cdot)$ est bien bijective. Par conséquent, $\gamma_s^{(t)}$ vit dans un sous espace vectoriel de $\mathbb{R}^Q$ de dimension $Q-1$. Le LDM introduit ci-dessous est donc définit sur cet espace :

$$\gamma_{s\setminus Q}^{(t)} | B, \nu^{(t)}, \Sigma \sim \mathcal{N}(B\nu^{(t)}, \Sigma),$$

où $\gamma_{s\setminus Q}^{(t)}$ désigne le vecteur $\gamma_s^{(t)}$ privé de sa dernière composante. Les matrices Σ et B sont de taille $(Q-1)\times(Q-1)$ alors que $\nu^{(t)}$ est un vecteur réel à $Q-1$ composantes. D'un point de vue modèle, les vecteurs $\nu^{(t)}$ caractérisent le mécanisme d'évolution global des proportions au cours du temps. C'est à partir de cette tendance que se construisent les différences de dynamique entre les sous-graphes. Le reste de dRSM fait intervenir une couche cachée standard pour les LDM :

$$\begin{cases} \nu^{(t)} = A\nu^{(t-1)} + \omega \\ \nu^{(1)} = u_0 + u, \end{cases}$$

où les termes de bruit (blanc) sont supposés Gaussien et indépendant :

$$\begin{cases} \omega | \Phi \sim \mathcal{N}(0, \Phi) \\ u | V_0 \sim \mathcal{N}(0, V_0). \end{cases}$$

À nouveau, A, ϕ, et V_0 sont des matrices de taille $(Q-1)\times(Q-1)$ et u_0 est un vecteur réel à $Q-1$ composantes. Les modèles basés sur LDM souffrent de problèmes importants d'identifiabilité. Plusieurs choix sont possibles pour régler cette question donc l'inférence est présentée ci-dessous dans un cadre général. En pratique, dans toutes les expériences que nous avons réalisées, nous avons fixé $A = B = V_0 = I_{Q-1}$ et toutes les composantes de u_0 à zéro. Nous notons ici $\theta = (u_0, A, B, \Phi, V_0, \Sigma, \mu)$ l'ensemble des paramètres et $\nu = (\nu^{(t)})_t$, $\gamma = (\gamma_s^{(t)})_{st}$, et $Z = (Z_{ik}^{(t)})_{ikt}$ l'ensemble des variables latentes du modèle.

3.2.1.b Inférence

L'inférence de dRSM sur données réelles pose des problèmes similaires à ceux rencontrés pour le modèle OSBM autorisant la recherche de clusters chevauchants et présentés en section 2.3. Tout d'abord, le modèle graphique impose que la log-vraisemblance a un coût algorithmique d'évaluation prohibitif et que la loi a posteriori des variables cachées n'a pas de forme analytique. De plus, une fonction $f(\cdot)$ non linéaire apparaissant dans le modèle, un algorithme standard de type VEM ne peut être dérivé directement. À l'image de nos travaux pour OSBM, nous avons donc construit une borne inférieure sur chaque terme de normalisation $C(\gamma_s^{(t)}) = \exp(\sum_{\ell=1}^{Q} \gamma_{s\ell}^{(t)})$. Les bornes associées dépendent alors de paramètres variationnels $\xi_s^{(t)} \in \mathbb{R}^{*+}$, fonctions des sous-graphes et du temps.

Proposition 3.1. *Pour tout ensemble $\xi = (\xi_s^{(t)})$ où $\xi_s^{(t)} \in \mathbb{R}^{*+}$ et toute loi $r(Z, \gamma, \nu)$, une borne inférieure de la log-vraisemblance est donnée par :*

$$\begin{aligned} &\mathcal{L}_Q(r; \theta, \xi) \\ &= \sum_Z \int\int r(Z, \gamma, \nu) \log \frac{p(X|Z, \mu, Q) h_Q(Z, \gamma, \xi) p(\gamma_{\setminus Q} | B, \nu, \Sigma, Q) p(\nu | u_0, A, \Phi, V_0, Q)}{r(Z, \gamma, \nu)} d\gamma d\nu, \end{aligned} \tag{3.2}$$

où

$$\log h_Q(Z,\gamma,\xi) = \sum_{t=1}^{T}\sum_{q=1}^{Q}\sum_{i=1}^{n}\sum_{s=1}^{S} y_{is} Z_{iq}^{(t)} \Big(\gamma_{sq}^{(t)} - \big(\xi_s^{-1(t)} \sum_{r=1}^{Q} \exp(\gamma_{sr}^{(t)}) - 1 + \log(\xi_s^{(t)})\big)\Big).$$

L'optimisation fonctionnelle de $\mathcal{L}(r;\theta,\xi)$ nécessite d'imposer des contraintes sur r. Comme précédemment, un espace de lois factorisables a été considéré :

$$r(Z,\gamma,\nu) = r(Z)r(\gamma)r(\nu) = \Big(\prod_{t=1}^{T}\prod_{i=1}^{n} r(Z_i^{(t)})\Big) r(\gamma)r(\nu).$$

De plus, une forme paramétrique a été choisie pour $r(\gamma)$:

$$r(\gamma) = \prod_{t=1}^{T}\prod_{s=1}^{S}\prod_{q=1}^{Q} \mathcal{N}(\gamma_{sq}^{(t)}; \hat{\gamma}_{sq}^{(t)}, \hat{\sigma}_{sq}^{2^{(t)}}).$$

Afin de respecter la contrainte de bijectivité de $f(\cdot)$, tous les termes $(\hat{\gamma}_{sQ}^{(t)}, \hat{\sigma}_{sQ}^{2^{(t)}})$ sont fixés à 0 de manière à imposer une masse de Dirac en 0 pour la dernière composante des vecteurs $\gamma_s^{(t)}$.

Un algorithme d'optimisation similaire à celui introduit pour OSBM en section 2.3.3.b est alors considéré pour l'inférence. Ainsi, la borne inférieure est maximisée par rapport à r, θ, et ξ de manière alternée jusqu'à convergence. L'optimisation des termes dans ξ permet notamment d'améliorer les approximations locales des constantes de normalisation. Un point notable concerne la loi $r(\gamma)$. À l'optimum, nous avons montré que sa forme fonctionnelle était celle d'un LDM particulier avec pour observations les vecteurs $x^{(t)} = \sum_{s=1}^{S} \hat{\gamma}_s^{(t)}/S$.

Proposition 3.2. *La loi $r(\gamma)$ maximisant la borne inférieure* (3.2) *est donnée par :*

$$r(\nu) \propto p(\nu^{(1)}|u_0,V_0)\Big[\prod_{t=2}^{T} p(\nu^{(t)}|\nu^{(t-1)},A,\Phi)\Big]\Big[\prod_{t=1}^{T} \mathcal{N}(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}; B\nu^{(t)}, \frac{\Sigma}{S})\Big].$$

Les paramètres dans $\theta' = (u_0, A, B, \Phi, V_0, \Sigma/S)$ peuvent donc être estimés directement à partir du filtre de Kalman et des équations de Rauch-Tung-Striebel (RTS) (Rauch, Tung et Striebel, 1965) appliquées aux observations $x^{(t)}$. De plus, les espérances et matrices de covariance des $\nu^{(t)}$ sachant les données sont déterminées à l'aide de récursions de type backward-forward.

Concernant la sélection de modèles, nous avons eu recourt à un critère de type BIC où la vraisemblance des données, qui n'est malheureusement pas calculable, est remplacée par son approximation variationnelle $\mathcal{L}(r;\theta,\xi)$, à convergence.

3.2.1.c Expérimentations numériques

Plusieurs études sur simulations numériques ont été réalisées dans Zreik, Latouche et Bouveyron, 2016 afin de tester la pertinence de la méthodologie d'inférence proposée. Nous présentons ici les résultats d'évaluation de la capacité de l'algorithme à retrouver les clusters, lorsque Q est donné. Cinq scénarios ont ainsi été considérés comme indiqué dans le tableau 3.1. Le type $c = 0$ correspond à l'absence d'arc. $Q = 4$ clusters apparaissent alors dans les réseaux avec les probabilités intra cluster $1 - \mu_{qq0}$ suivante : 0.9, 0.6, 0.5, 0.4. De

Paramètres	Scénario 0	Scénario 1	Scénario 2	Scénario 3	Scénario 4
n	300				
Q	4				
T	10 (indep.)	10 (LDM)			
S	1	1	1	2	2
C	1	1	1	1	2
$(\mu_{qq0})_{q=1,..,Q}$	(0.1,0.4,0.5,0.6)				
$\mu_{qr0,q\neq r}$	0.99		0.8	0.99	
$\mu_{qrc,c\neq 0}$	$(1-\mu_{qr0})/C$				

Table 3.1 Paramètres des cinq scénarios utilisés pour les expériences sur simulations numériques. Dans le Scénario 0, les réseaux sont générés sans dépendance temporelle. Dans tous les autres scénarios, un LDM est considéré.

Méthode	Scénario 0	Scénario 1	Scénario 2	Scénario 3	Scénario 4
SBM	0.10±0.04	0.12±0.05	0.18±0.07	0.14±0.09	–
RSM	–	–	–	–	0.01±0.01
dM3SBM	0.36±0.09	0.30±0.16	0.25±0.16	0.32±0.20	–
dRSM	1.00±0.00	0.98±0.04	0.90±0.20	0.97±0.07	0.75±0.24

Table 3.2 Résultats de clustering pour quatre méthodes appliquées à des réseaux simulés selon les cinq scénarios considérés. Les moyennes d'ARI et les écarts types sont indiqués pour les 20 réseaux générés dans chaque scénario.

plus, de faibles probabilités inter cluster $1-\mu_{qr0,q\neq r}$ sont fixées : 0.01 ou 0.2, suivant les scénarios. Les réseaux générés suivant tous ces schémas de simulation sont donc construits à partir de communautés où les probabilités de connexion sont plus fortes entre des nœuds de même cluster. Enfin, les différents types $c \neq 0$ ont la même probabilité d'apparition. Dans le scénario 0, il n'y a pas de dépendance temporelle et les matrices d'adjacence $X^{(t)}$ sont générées de manière indépendante. En revanche, dans tous les autres scénarios, un LDM est utilisé. Les paramètres A, B, et V_0 sont fixés à I_{Q-1}. De plus, nous avons $\Sigma = 0.1 \times I_{Q-1}$, $\Phi = 0.01 \times I_{Q-1}$ et toutes les composantes de u_0 à 0. Pour $S > 1$, les nœuds sont répartis de manière uniforme dans les sous-graphes.

L'algorithme d'inférence associé à dRSM a été comparé avec les algorithmes VBEM pour SBM et RSM présentés en section 2.2.1 et 2.4 respectivement, ainsi qu'avec une autre méthode appelée dM3SBM (Ho, Song et Xing, 2011) permettant l'analyse de réseaux dynamiques où les clusters sont construits dans des sous-graphes. SBM ne pouvant modéliser que des connexions binaires, les arcs sont binarisés pour l'étude correspondante. Finalement, pour chaque scénario, 20 réseaux sont simulés et analysés par les différentes méthodes. La qualité de l'estimation des clusters est évaluée à l'aide d'un critère ARI. Les résultats, présentés dans le tableau 3.2, sont sensiblement meilleurs pour dRSM ayant obtenue des ARI proches de 1.

3.2.2 Construction d'une partition du temps

Nous présentons dans cette section les travaux publiés dans Corneli, Latouche et Rossi, 2016b dans un cas orienté. Ils peuvent être vus comme une extension des travaux introduits en section 2.2.2, au cas des réseaux dynamiques. Contrairement à la section précédente, les nœuds ne peuvent pas ici changer de clusters au cours du temps. En effet, comme pour le modèle SBM standard, chaque nœud i est associé à un vecteur binaire Z_i, non indicé par le

temps, tiré à partir d'une loi multinomiale. Ainsi, le nœud i appartient au cluster de nœuds q avec probabilité π_q et le vecteur $\pi = (\pi_q)_q$ désigne les poids de tous les Q clusters.

En revanche, les paramètres de connectivité sont autorisés à évoluer au cours du temps entre les paires de clusters. Afin de limiter le nombre de ces paramètres et de faciliter l'interprétation des résultats, l'objectif est également de construire une partition des intervalles de temps. Ainsi, chaque intervalle de temps I_u est associé à un cluster temporel parmi D, grâce au vecteur binaire Y_u :

$$Y_u|\beta \sim \mathcal{M}(1, \beta),$$

où $\beta = (\beta_d)_d$. Le nombre d'interactions $X_{ij}^{(u)}$ ayant lieu entre i et j durant I_u est alors généré à partir d'une loi de Poisson conditionnelle :

$$X_{ij}^{(u)}|Z_{iq}Z_{jr}Y_{ud} = 1, \lambda \sim \mathcal{P}(\lambda_{qrd}),$$

où $\lambda = (\lambda_{qrd})_{qrd}$. Le modèle associé est appelé d1SBM dans ce mémoire. L'ensemble des paramètres de d1SBM est noté $\theta = (\pi, \beta, \lambda)$.

3.2.2.a Inférence

Pour réaliser l'inférence de d1SBM sur données réelles, nous nous sommes inspirés de l'algorithme glouton de Côme et Latouche, 2015, afin de se concentrer sur la tache de clustering. Une loi a priori entièrement factorisée sur l'ensemble des paramètres est donc introduite :

$$p(\theta) = p(\pi)p(\beta)p(\lambda).$$

Des lois a priori de Dirichlet sont alors considérées pour les vecteurs de proportions de clusters :

$$p(\pi) = \mathrm{Dir}(\pi; \alpha, \ldots, \alpha),$$

et

$$p(\beta) = \mathrm{Dir}(\beta; \gamma, \ldots, \gamma).$$

Pour une discussion sur le choix des hyperparamètres, nous renvoyons le lecteur à la section 2.2. Dans toutes les expériences, ils ont été fixés à 1 afin d'obtenir des lois uniformes sur les simplexes correspondant. De plus, tous les paramètres d'intensité dans λ sont supposés *iid* a priori et une loi Gamma est utilisée pour caractériser λ_{qrd} :

$$p(\lambda_{qrd}) = \mathrm{Gam}(\lambda_{qrd}; a, b).$$

Cette loi devient non informative que dans un cas limite, lorsque b tend vers 0. Dans nos expériences, nous avons fixé $a = b = 1$. Dans ce cadre, il n'est pas nécessaire d'avoir recours à des approximations asymptotiques de type Laplace ou Stirling. En effet, les choix fait concernant la construction du modèle probabiliste et des lois a priori permettent d'obtenir une expression exacte de la vraisemblance intégrée des données complétées.

Proposition 3.3. *La vraisemblance intégrée des données complétées de d1SBM s'écrit explicitement :*

$$\begin{aligned} &p(X, Z, Y|Q, D) \\ &= \left(\prod_{q,r}^{Q} \prod_{d=1}^{D} \frac{b^a}{\Gamma(a) P_{qrd}} \frac{\Gamma(S_{qrd} + a)}{(R_{qrd} + b)^{S_{qrd}+a}} \right) \frac{\Gamma(\alpha Q)}{\Gamma(\alpha)^Q} \frac{\prod_{q=1}^{Q} \Gamma(|A_q| + \alpha)}{\Gamma(n + \alpha Q)} \frac{\Gamma(\gamma D)}{\Gamma(\gamma)^D} \frac{\prod_{d=1}^{D} \Gamma(|C_d| + \gamma)}{\Gamma(U + \gamma D)}, \end{aligned} \tag{3.3}$$

où

- $|A_q| = \sum_{i=1}^{n} Z_{iq}$;
- $|C_d| = \sum_{u=1}^{U} Y_{ud}$;
- $S_{qrd} = \sum_{i \neq j}^{n} \sum_{u=1}^{U} Z_{iq} Z_{jr} Y_{ud} X_{ij}^{(u)}$;
- $P_{qrd} = \prod_{i \neq j}^{n} \prod_{u=1}^{U} (X_{ij}^{(u)}!)^{Z_{iq} Z_{jr} Y_{ud}}$;
- $R_{qrd} = |A_q||A_r||C_d|$ *si* $q \neq r$, $|A_q|(|A_q| - 1)|C_d|$ *sinon.*

Un critère ICL exact correspondant au log de (3.3) est donc considéré pour l'inférence. Le critère étant auto-pénalisant et dépendant (Z, Y, Q, D), sa maximisation par rapport à tous ces termes permet simultanément d'estimer le nombre de clusters de nœuds Q, le nombre de clusters d'intervalles de temps D, et les partitions associées. Malheureusement, ce problème d'optimisation est combinatoire et par conséquent nous avons dérivé un algorithme glouton afin de construire un optimum local. Ainsi, à l'image de la procédure de maximisation introduite en section 2.2.2, les nœuds et les intervalles de temps sont initialement placés dans des clusters. Des mouvements d'échange entre des clusters et des fusions sont alors appliqués de manière itérée, s'ils induisent une augmentation de l'ICL. Ces mouvements sont testés à la fois pour les nœuds et les intervalles de temps. Dans le cas où plusieurs mouvements sont possibles, celui induisant l'accroissement maximal de l'ICL est retenu. Ces étapes sont répétées jusqu'à convergence du critère. Pour plus de détails, nous renvoyons le lecteur à la section 2.2.2.

3.2.2.b Expérimentations numériques

Dans cette section, nous présentons une partie des expériences que nous avons réalisées sur données simulées pour tester l'algorithme glouton d'inférence. Uniquement les résultats obtenus sur des réseaux construits à partir de communautés sont indiqués. Ainsi, n est fixé à 50 et Q à 3. Les interactions ont lieu sur $U = 50$ intervalles de temps appartenant à $D = 3$ clusters de temps. De plus, les poids des clusters sont fixés à $\pi = \beta = (1/3, 1/3, 1/3)$. Finalement, les termes λ_{qrd} sont définis à partir de la fonction $\Lambda(\cdot)$:

$$\Lambda(u) = L\mathbb{1}_{C_1}(u) + \sqrt{\gamma} L\mathbb{1}_{C_2}(u) + \gamma L\mathbb{1}_{C_3}(u), \quad u \in \{1, \ldots, 50\},$$

où

$$L = \begin{pmatrix} \psi & 2 & 2 \\ 2 & \psi & 2 \\ 2 & 2 & \psi \end{pmatrix},$$

et $\mathbb{1}_{C_d}(u) = 1$ si l'intervalle I_u fait parti du cluster de temps d. Ainsi, dans le cas où $\psi > 2$, L permet de générer des réseaux où les nœuds de même cluster se connectent préférentiellement. Le niveau global d'intensité dans L est lui multiplié par un terme dépendant du temps, 1, $\sqrt{\gamma}$, ou γ, suivant les cas. Les paramètres libres sont donc ψ et γ. Une grille de valeurs a alors été constituée pour ψ ainsi que pour γ, et pour chaque paire (ψ, γ), 50 réseaux dynamiques ont été simulés selon d1SBM. L'algorithme glouton a été appliqué sur chaque réseau et les résultats évalués grâce au critère ARI. Ce dernier permet ici de résumer la qualité de l'estimation de (Z, Y, Q, D), tous ces termes étant inférés par la méthodologie.

Les valeurs d'ARI sont présentées sous la forme de boîtes à moustaches pour le clustering des intervalles de temps à $\psi = 2$ dans la figure 3.1 et pour le clustering des nœuds à $\gamma = 1$ en figure 3.2. Les scénarios associés sont critiques puisqu'à $\psi = 2$, les niveaux d'intensité intra et inter clusters sont identiques, il n'y a donc pas de cluster de nœuds. De manière similaire,

$\gamma = 1$ induit des réseaux sans structure temporelle. Dans le premier cas, nous observons qu'à mesure que la valeur de γ augmente, la méthodologie tend à retrouver parfaitement les clusters temporels. Les résultats sont en accord avec ce constat, dans le second cas. En effet, lorsque la valeur de ψ augmente, les clusters de nœuds sont bien identifiés.

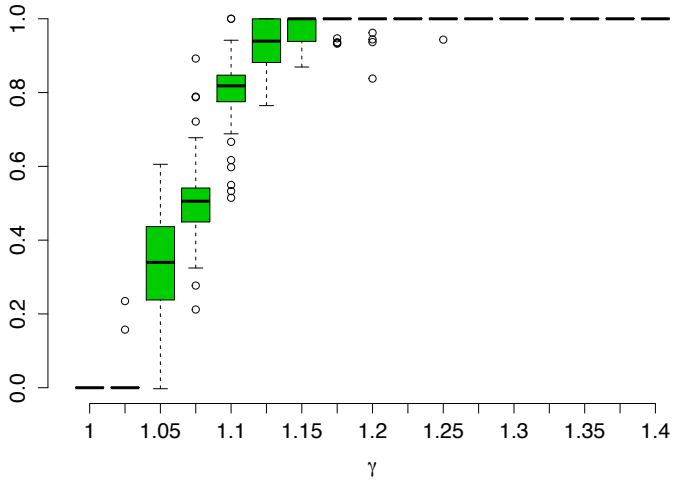


Figure 3.1 – Boîte à moustaches des résultats d'ARI en fonction de γ pour le clustering des intervalles de temps. Le paramètre ψ est fixé à 2.

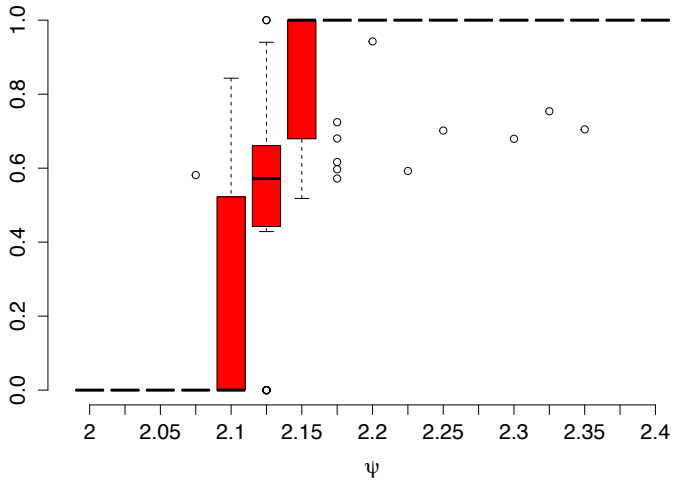


Figure 3.2 – Boîte à moustaches des résultats d'ARI en fonction de ψ pour le clustering des nœuds. Le paramètre γ est fixé à 1.

3.3 Temps continu et segmentation

Comme mentionné précédemment, les méthodes développées dans la littérature pour l'analyse des réseaux dynamiques fonctionnent généralement à temps discret. Dans certaines bases de données, les réseaux sont stockés sous la forme d'une série d'images indicées par

le temps. Ils peuvent donc directement être étudiés à partir de ces outils. Lorsque ce n'est pas le cas, un pré-traitement des données est nécessaire, impliquant de définir des intervalles de temps afin d'agréger les interactions observées. Le choix des intervalles peut alors avoir un impact fort sur les résultats d'analyse. Ce point clé a été au coeur de plusieurs travaux publiés récemment dans la littérature. Il est par exemple discuté dans Matias, Rebafka et Villers, 2015. Une partie de la communauté statistique, travaillant sur la modélisation des réseaux, s'est donc tournée vers la prise en compte d'un temps continu, abandonnant ainsi la construction des images. Sur cette thématique, nous avons pris un point de vue intermédiaire dans Corneli, Latouche et Rossi, à paraître. En effet, à l'aide de processus ponctuelles, nous avons bien autorisé l'extraction d'informations directement à partir de l'ensemble $\mathcal{E}$ des interactions. En complément, grâce à des modèles utilisés pour la segmentation, nous avons également permis la détection de segments, et donc d'images, sur lesquels les intensités d'interaction sont homogènes. Ces images aident à l'interprétation des résultats.

3.3.1 Processus ponctuels et comptages

Nous reprenons ici le cadre et les notations introduites dans l'introduction de ce chapitre. Comme pour le modèle SBM standard, chaque nœud i du réseau est alloué à un cluster q ne dépendant pas du temps, avec probabilité π_q :

$$Z_i|\pi \sim \mathcal{M}(1,\pi),$$

où $\pi = (\pi_1, \ldots, \pi_Q)$. Notons ensuite que les temps d'interaction entre i et j dans $\mathcal{A}^{(i,j)}$ peuvent être vus comme la réalisation d'un processus ponctuel. En tant que tel, le processus prend ses valeurs sur $[0,T]$ et est naturellement associé à un processus de comptage que nous désignons par $\{M^{(i,j)}(t)\}_{t\in[0,T]}$. La variable aléatoire $M^{(i,j)}(t)$ compte alors le nombre d'interactions ayant lieu entre i et j avant (ou exactement) le temps t :

$$M^{(i,j)}(t) = \left|\mathcal{A}^{(i,j)} \cap]0,t]\right|.$$

Afin de modéliser les temps d'interaction, nous avons considéré un processus ponctuel de Poisson non homogène, non homogeneous Poisson point process (NHPPP) en anglais. Ce dernier est caractérisé par une fonction d'intensité $\kappa^{(i,j)}(\cdot)$, positive et intégrable sur $[0,T]$. Par conséquent, en notant :

$$\bar{\kappa}^{(i,j)}(t) = \int_0^t \kappa^{(i,j)}(s)ds, \qquad t \leq T,$$

$M^{(i,j)}(t)$ suit une loi de Poisson de paramètre $\bar{\kappa}^{(i,j)}(t)$. Le réseau étant ici non orienté, $n(n-1)/2$ NHPPP sont utilisés. Comme pour le modèle SBM, l'information de connexion est portée par les clusters. Nous présentons ici la méthodologie dans le cas de réseaux non orientés donc $Q(Q+1)/2$ fonctions $\lambda = \{\lambda_{qr}(\cdot)\}_{qr}$ positives intégrables sur $[0,T]$ sont employées telles que $\kappa^{(i,j)}(t) = \lambda_{Z_i,Z_j}(t), \forall t \in [0,T]$. Le modèle ainsi obtenu est appelé *d2SBM*.

Proposition 3.4. *La vraisemblance des données complétées du modèle d2SBM est donnée par :*

$$p(\mathcal{E}, Z|\lambda,\pi,Q) = \exp\left(-\sum_{i<j}^{n} \Lambda_{Z_iZ_j}(T)\right) \prod_{m=1}^{M} \lambda_{Z_{i_m}Z_{j_m}}(\nu_m) \prod_{i=1}^{n} \pi_{Z_i},$$

où

$$\Lambda_{qr}(t) = \int_0^t \lambda_{qr}(s)ds, \qquad t \leq T.$$

3.3.2 Points de rupture

Les fonctions d'intensité des NHPPP sont supposées constantes par morceau. Plus précisément, nous faisons l'hypothèse que les D intervalles, ou segments, sur lesquels elles sont constantes sont partagés par toutes les fonctions. En d'autres termes, $D-1$ points de rupture sont utilisés :

$$0 = \eta_0 < \eta_1 < \cdots < \eta_{D-1} < \eta_D = T,$$

et

$$\lambda_{qr}(t) = \sum_{d=1}^{D} \lambda_{qrd} \mathbb{1}_{[\eta_{d-1}\eta_d[}(t), \qquad \forall q, r \in \{1, \ldots, Q\}.$$

Une conséquence cruciale de cette hypothèse est que sur l'intervalle $[\eta_d, \eta_{d+1}[$, tous les processus ponctuels de Poisson sont homogènes. En revanche, les interactions peuvent changer de segments de manière arbitraire, autorisant ainsi des changements radicaux dans les niveaux d'intensité entre les paires de clusters. Enfin, l'hypothèse imposant une contrainte sur les données, une forme plus simple est obtenue pour la vraisemblance des données complétées.

Proposition 3.5. *Sous l'hypothèse que les fonctions d'intensité des NHPPP sont toutes constantes sur D intervalles, la log-vraisemblance des données complétées du modèle d2SBM s'écrit :*

$$\log p(\mathcal{E}, Z|\theta, Q, D) = \\ -\sum_{d=1}^{D}\sum_{q,r}^{Q}\left[\lambda_{qrd}\Delta_d\left(\sum_{i<j}^{n} Z_{iq}Z_{jr}\right) - \log(\lambda_{qrd})\left(\sum_{i<j}^{n} Z_{iq}Z_{jr}X_{ij}^{(d)}\right)\right] + \sum_{i=1}^{n}\sum_{q=1}^{Q} Z_{iq}\log\pi_q,$$

où $\theta = (\pi, \eta, \lambda)$ avec $\eta = (\eta_d)_d$ et Δ_d est la taille de l'intervalle $[\eta_{d-1}, \eta_d[$. Le terme $X_{ij}^{(d)}$ désigne ici le nombre d'interactions ayant lieu entre i et j, sur ce segment. Contrairement à l'algorithme glouton présenté en section 3.2.2, où une décomposition de $[0, T]$ en intervalles I_u prédéfinis est utilisée, il est important de noter que les segments eux même sont maintenant à estimer à partir des données.

3.3.3 Vraisemblance pénalisée et approximations variationnelles

L'inférence de d2SBM implique l'estimation de Q, D, Z, et θ. Notons en particulier que l'estimation de D et η est liée à un problème de segmentation, c'est-à-dire d'optimisation sur un espace incluant le nombre de segments et la position des points de rupture. De nombreuses approches ont été proposées dans la littérature pour ce type de tache. Dans Corneli, Latouche et Rossi, à paraître, nous avons choisi de nous appuyer sur la méthode exacte de découpage en temps linéaire, pruned exact linear time (PELT) method en anglais, de Killick, Fearnhead et Eckley, 2012. Sous certaines conditions, discutées plus bas concernant la fonction de gain définie sur les segments, une solution optimale au problème de segmentation peut être obtenue, tout en réduisant l'espace d'exploration au fur et à mesure des itérations. Ces réductions successives permettent de diminuer considérablement le temps d'exploration. Afin de construire ce type de fonction de gain, un cadre particulier d'inférence a été introduit.

Tout d'abord, dans le but de pénaliser le nombre de clusters Q ainsi que le nombre de segments D, la log-vraisemblance intégrée par rapport à π et λ est considérée :

$$\log p(\mathcal{E}|Q, \eta, D) = \log\left(\int_{\lambda,\pi} p(\mathcal{E}, \lambda, \pi|Q, \eta, D)d\lambda d\pi\right).$$

Suite à la marginalisation, cette dernière est uniquement fonction de Q, η, et D. Malheureusement, n'ayant pas de forme analytique, nous avons proposé de la remplacer par une approximation de type BIC :

$$\log \tilde{p}(\mathcal{E}|Q,\eta,D) = \max_{\lambda,\pi} \log p(\mathcal{E}|Q,\eta,D,\lambda,\pi) - \frac{1}{2}C(Q,D)\log\alpha, \tag{3.4}$$

où

$$C(Q,D) = Q - 1 + \frac{Q(Q+1)D}{2},$$

est le nombre de paramètres de d2SBM. Le terme α est lié au nombre d'observations présentes dans les données et est discuté plus bas. Le premier élément de la log-vraisemblance pénalisée (3.4) est ici évalué au maximum en (λ,π). Le problème d'inférence s'écrit alors :

$$\max_{Q,\eta,D,\lambda,\pi} \log p(\mathcal{E}|Q,\eta,D,\lambda,\pi) - \frac{1}{2}C(Q,D)\log\alpha.$$

Le modèle d2SBM dépend d'une couche cachée pour le clustering des nœuds et codée par la matrice Z. Le terme $\log p(\mathcal{E}|Q,\eta,D,\lambda,\pi)$ fait donc intervenir une marginalisation sur toutes les Q^n matrices Z possibles. Il ne peut donc pas être calculé en pratique. Par conséquent, il est remplacé par une approximation variationnelle.

Proposition 3.6. *Pour toute loi $r(Z) = \prod_{i=1}^{n}\prod_{q=1}^{Q}\tau_{iq}^{Z_{iq}}$ factorisable, une borne inférieure de $\log p(\mathcal{E}|Q,\eta,D,\lambda,\pi) - \frac{1}{2}C(Q,D)\log\alpha$ est donnée par :*

$$\begin{aligned} &f(r,Q,\eta,D,\lambda,\pi) \\ &\quad = -\sum_{d=1}^{D}\sum_{q,r}^{Q}\left[\lambda_{qrd}\Delta_d\left(\sum_{i<j}^{n}\tau_{iq}\tau_{jr}\right) - \log(\lambda_{qrd})\left(\sum_{i<j}^{n}\tau_{iq}\tau_{jr}X_{ij}^{(d)}\right)\right] \\ &\qquad\qquad + \sum_{i=1}^{n}\sum_{q=1}^{Q}\tau_{iq}\log\frac{\pi_q}{\tau_{iq}} - \frac{1}{2}C(Q,D)\log\alpha. \end{aligned}$$

Comme précédemment, τ_{iq} est l'approximation variationnelle de la probabilité que le nœud i soit dans le cluster q, sachant les données. Le problème d'optimisation devient donc :

$$\max_{Q,\eta,D,\lambda,\pi,r} f(r,Q,\eta,D,\lambda,\pi).$$

3.3.4 Optimisation

L'optimisation de $f(\cdot)$ par rapport à r, π, et λ ne pose pas de problème particulier.

Proposition 3.7. *Les estimateurs de r, π, et λ maximisant $f(r,Q,\eta,D,\lambda,\pi)$ sont donnés par :*

$r(Z) = \prod_{i=1}^{n}\prod_{q=1}^{Q}\tau_{iq}^{Z_{iq}}$ **avec :**

$$\tau_{iq} \propto \pi_q \exp\left\{-\sum_{d=1}^{D}\sum_{r=1}^{Q}\left[\lambda_{qrd}\Delta_d\left(\sum_{i\neq j}^{n}\tau_{jr}\right) - \log(\lambda_{qrd})\left(\sum_{i\neq j}^{n}\tau_{jr}X_{ij}^{(d)}\right)\right]\right\},$$

et $\sum_{q=1}^{Q}\tau_{iq} = 1$.

$\hat{\pi}_q$:

$$\hat{\pi}_q = \frac{\sum_{i=1}^{n} \tau_{iq}}{n}.$$

$\hat{\lambda}_{qrd}$:

$$\hat{\lambda}_{qrd} = \begin{cases} \dfrac{\sum_{i<j}^{n} \tau_{iq}\tau_{jr} X_{ij}^{(d)}}{\Delta_d \sum_{i<j}^{n} \tau_{iq}\tau_{jr}} & si\ r > q, \\ \dfrac{\sum_{i\neq j}^{n} \tau_{iq}\tau_{jr} X_{ij}^{(d)}}{\Delta_d \sum_{i\neq j}^{n} \tau_{iq}\tau_{jr}} & si\ q = r. \end{cases}$$

Rappelant que

$$\max_{\eta,D,\lambda,\pi} f\left(r, Q, \eta, D, \lambda, \pi\right) = \max_{\eta,D} \max_{\lambda,\pi} f\left(r, Q, \eta, D, \lambda, \pi\right),$$

et à r et Q fixes, $f(\cdot)$ peut être évaluée en $\hat{\lambda}$ et $\hat{\pi}$ donnés en proposition 3.7. Elle peut alors s'écrire sous la forme :

$$\max_{\lambda,\pi} f\left(q(Z), Q, \eta, D, \lambda, \pi\right) = \sum_{d=1}^{D} \mathcal{G}([\eta_{d-1}, \eta_d[) - \frac{1}{2}\frac{Q(Q+1)D}{2}\log\alpha + const,$$

où tous les termes ne dépendant pas de η et D sont absorbés dans la constante *const* et

$$\mathcal{G}([\eta_{d-1}, \eta_d[) = -\sum_{q,r}^{Q} \left[\hat{\lambda}_{qrd}\Delta_d \left(\sum_{i<j}^{n} \tau_{iq}\tau_{jr} \right) - \log(\hat{\lambda}_{qrd}) \left(\sum_{i<j}^{n} \tau_{iq}\tau_{jr} X_{ij}^{(d)} \right) \right].$$

La fonction $f(\cdot)$, en $\hat{\lambda}$ et $\hat{\pi}$, est donc définie comme une somme de fonction de gains sur les segments. Cette forme permet d'utiliser la programmation dynamique pour parcourir tout l'espace des segments et trouver un estimateur global pour D ainsi que pour η, à (r, Q) fixes. Nous renvoyons le lecteur à la preuve associée dans Corneli, Latouche et Rossi, à paraître pour plus de détails.

La proposition suivante indique également que la méthode de découpage de PELT peut être utilisée afin de réduire le temps d'exploration des segments.

Proposition 3.8. *La condition de* Killick, Fearnhead *et* Eckley, *2012 permettant de réduire l'espace d'exploration lors de l'optimisation dynamique est vérifiée pour les fonctions gains construites dans ce cadre d'inférence et pour les approximations utilisées :*

$$\mathcal{G}([t_{u'}, t_u[) + \mathcal{G}([t_u, t_{u''}[) \geq \mathcal{G}([t_{u'}, t_{u''}[), \qquad \forall t_{u'} < t_u < t_{u''}.$$

3.3.5 Inférence

Comme montré dans Corneli, Latouche et Rossi, à paraître, un espace fini doit être considéré pour la segmentation et la recherche des points de rupture. Une grille a priori est donc introduite :

$$\mathcal{P} = \{(t_0, \ldots, t_U) | 0 = t_0 < t_1 < \cdots < t_U = T, \quad U \in \mathbb{N}^*\}.$$

La valeur maximale pour D est alors U et le terme α dans $f(\cdot)$ est égal à $Un(n-1)/2$. U est un paramètre de précision. En effet, une valeur importante pour U permet d'identifier les points de rupture sur la grille discrète, avec une plus grande précision, et inversement. Cette recherche a un coût algorithmique qui est directement impacté par U. Notons que, contrairement aux modèles temporels discrets décrits dans les sections précédentes de ce chapitre, le cadre de modélisation et d'inférence proposé ici offre un choix naturel pour la construction de la grille. La précision maximale est en effet obtenue en choisissant l'ensemble des interactions dans $\mathcal{E}$ pour $\mathcal{P}$. De cette grille prédéfinie sont construits les segments sur lesquels les niveaux d'intensité sont homogènes.

A Q fixe, les étapes de segmentation et d'optimisation par rapport à r sont donc alternées jusqu'à convergence de $f(\cdot)$. Rappelons que la segmentation optimise $f(\cdot)$ par rapport à η et D, λ et π étant remplacés par leur estimateur. A r et Q donnés, cette étape construit des optimums globaux de f. En revanche, l'inférence alternant entre la segmentation et l'optimisation en r, l'algorithme n'a pas de garantie de convergence vers un optimum global par rapport à r, η, D, λ, π. En pratique, le nombre de clusters de nœuds Q est inconnu. L'algorithme d'inférence est donc répété pour différentes valeurs de Q et la valeur maximisant $f(\cdot)$ est retenue comme estimateur.

3.3.6 Expérimentations numériques

Nous présentons ici le troisième scénario de simulations utilisé dans Corneli, Latouche et Rossi, à paraître pour comparer la méthodologie d'inférence pour d2SBM avec une méthode appelée MODL (Guigourès, Boullé et Rossi, 2015). Cette dernière permet également la recherche d'une partition des nœuds et une segmentation du temps. L'objectif de ce scénario est aussi de montrer que l'agrégation des interactions peut entrainer une perte importante d'informations. Ainsi, chaque réseau généré est constitué de 100 nœuds classés en deux groupes de 50 nœuds. L'intervalle $[0, 12]$ d'étude est alors coupé en quatre segments de même taille. De plus, les fonctions d'intensité sont fixées à :

$$\lambda_{Z_i Z_j}(t) = \begin{cases} 0.05\mathbb{1}_{I_1}(t) + 0.1\mathbb{1}_{I_2}(t) + 0.05\mathbb{1}_{I_3}(t) + 0.1\mathbb{1}_{I_4} & \text{si} \quad Z_i = Z_j \\ 0.1\mathbb{1}_{I_1}(t) + 0.05\mathbb{1}_{I_2}(t) + 0.1\mathbb{1}_{I_3}(t) + 0.05\mathbb{1}_{I_4} & \text{si} \quad Z_i \neq Z_j, \end{cases}$$

où I_d désigne le segment d. Par conséquent, les fonctions d'intensité intégrée sont toutes identiques :

$$\Lambda_{11}(T) = \Lambda_{12}(T) = \Lambda_{21}(T) = \Lambda_{22}(T) = 3.8.$$

Les clusters ne sont donc théoriquement pas distinguables si les interactions sont agrégées sur tout l'intervalle. 50 réseaux sont simulés suivant d2SBM pour ces paramètres et 10% des interactions observées sont modifiées uniformément au hasard (rewirering). Finalement, une méthode de type MCMC (SBM-Poisson) que nous avons introduite dans Nouedoui et Latouche, 2013, basée sur un processus de restaurant chinois adapté à un SBM avec connexions discrètes, est appliquée sur les données agrégées. Les résultats sont présentés en figure 3.3. Ils illustrent en particulier clairement que d2SBM produit des estimateurs plus pertinents que MODL, dans ce scénario. Nous observons également que SBM-Poisson est incapable d'identifier les clusters de nœuds à cause que l'agrégation des données.

Conclusion

Dans ce chapitre, nous avons étendu une partie des travaux présentés dans le chapitre précédent au cadre dynamique. Ainsi, à temps discret, nous avons d'abord dérivé le modèle dRSM

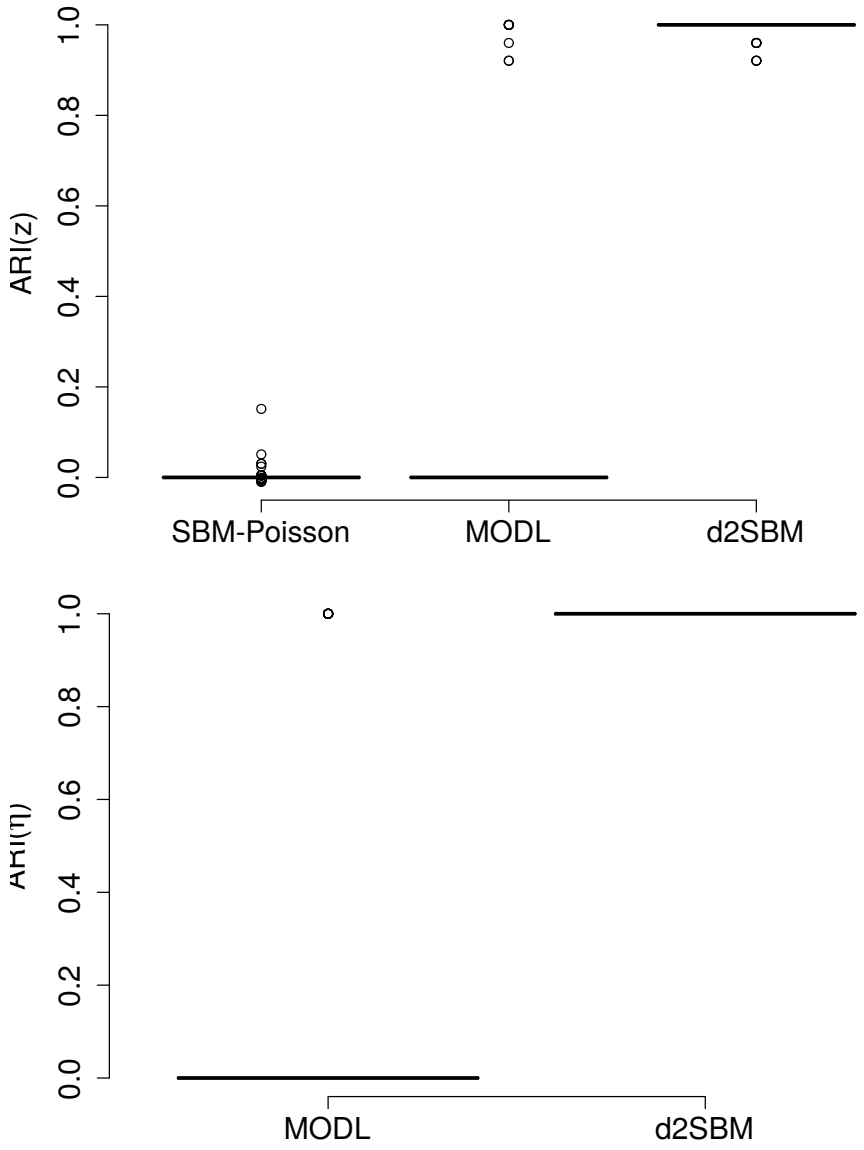


Figure 3.3 – Boîtes à moustaches des valeurs d'ARI trouvées sur données simulées. En haut : ARI pour le clustering des nœuds. En bas : ARI pour les points de rupture.

permettant d'intégrer une pré-classification des nœuds fournie par des experts. L'inférence est réalisée à l'aide d'une procédure de type backward-forward. Ensuite, un algorithme d'optimisation a été construit afin de chercher notamment des clusters d'intervalles de temps. Contrairement au modèle dRSM, les paramètres de connectivité peuvent changer au cours du temps, mais pas les clusters eux mêmes. Enfin, à temps continu, un cadre méthodologique a été adopté autorisant l'identification de ruptures dans les taux d'interactions entre les sommets du réseau. Les segments identifiés autorisent la construction d'images du réseau où les niveaux d'interactions sont homogènes.

4

De l'étude des graphons

4.1 Introduction . . . 50
4.2 Estimation dans le modèle W-graphe . . . 51
4.2.1 Estimation de la fonction graphon 52
4.2.2 Probabilités d'apparition de motifs 53
4.2.3 Expérimentations numériques 53
4.3 Qualité de l'ajustement : approche Bayésienne . . . 55
4.3.1 Régression logistique et structure résiduelle 55
4.3.2 Comparaison Bayésienne de modèles 55
4.3.3 Inférence 56
4.3.4 Expérimentations numériques 59
4.4 Qualité de l'ajustement : tests . . . 61
4.4.1 Modèle indépendant 61
4.4.2 Modèle de W-graphe 62
4.4.3 Expérimentations numériques 64

Cette thématique de recherche s'intéresse au modèle de W-graphe généralisant la plupart des modèles de graphes aléatoires existants. La recherche en statistique pour l'étude des réseaux s'est largement tournée vers ce modèle ces cinq dernières années. L'objectif premier est de trouver des stratégies d'inférence pour le W-graphe, sur données réelles. Mes contributions sont organisées ci-dessous en trois sous-axes. Les travaux du premier sous-axe, publiés dans un article de journal (Latouche et Robin, 2016), permettent l'estimation de la fonction graphon caractérisant le W-graphe. Le second sous-axe se concentre sur l'évaluation de la qualité d'ajustement lorsque la régression logistique est utilisée pour prédire les connexions d'un réseau à partir de covariables. Le modèle de W-graphe est alors utilisé pour décrire les structures cachées non expliquées. Ces travaux sont publiés dans un article de journal (Latouche, Robin et Ouadah, à paraître). Enfin, le troisième sous-axe présente les tests de Ouadah, Robin et Latouche, 2017 ainsi que les résultats asymptotiques associés. Ces

travaux permettent en pratique de tester l'adéquation entre un modèle de graphe aléatoire et un réseau donné. Uniquement des réseaux non orientés sont ici considérés.

4.1 Introduction

Les réseaux sont par construction des structures décrivant des connexions entre des objects d'intérêt. Théoriquement, la notion de connexion est à rapprocher de la notion de dépendance statistique. En effet, les modèles existants de graphe aléatoire font presque tous l'hypothèse de l'existence d'une couche cachée dans le modèle graphique. Conditionnellement aux variables latentes, les arcs ou arêtes sont alors indépendants. En revanche, marginalement, ces derniers sont tous dépendants entre eux. La couche cachée est donc porteuse d'une information clé que les techniques d'inférence cherchent à extraire sur données réelles. Ainsi, la notion d'indépendance directe entre les connexions est remplacée par la notion d'échangeabilité.

L'échangeabilité apparaît à l'origine dans le théorème de représentation de Finetti, 1931. Le théorème indique qu'une séquence infinie de variables aléatoires $X_1, \ldots$, où $X_i \in \{0, 1\}$, est échangeable si et seulement si il existe une loi $p(\theta)$ telle que les X_i sont iid conditionnellement à θ. Ce théorème est généralisé par Hewitt et Savage, 1955 pour une séquence de variables aléatoires à valeurs dans un ensemble $\mathbf{Z}$ quelconque. Dans le cas des réseaux, une matrice d'adjacence $X = (X_{ij})_{i,j\in\mathbb{N}}$ où $X_{ij} \in \mathbf{Z}$ de taille infinie est considérée. Le théorème de Aldous, 1981 ; Hoover, 1979 pose alors que X est échangeable si et seulement si il existe une fonction $f : [0, 1]^3 \to \mathbf{Z}$ telle que $X_{ij}|f, U_i, U_j, U_{ij} \sim f(U_i, U_j, U_{ij})$ où les variables aléatoires $(U_i)_{i\in\mathcal{N}}$ et $(U_{ij})_{i<j\in\mathcal{N}}$, avec $U_{ij} = U_{ji}$, sont iid tirées à partir d'une loi $\mathcal{U}([0, 1])$, uniforme continue sur $[0, 1]$. Dans le cas des réseaux non orientés binaires ($\mathbf{Z} = \{0, 1\}$), le théorème de représentation de Aldous-Hoover peut s'exprimer à l'aide d'une fonction appelée graphon $W : [0, 1]^2 \to [0, 1]$ telle que $X_{ij}|W, U_i, U_j \sim \mathcal{B}(W(U_i, U_j))$, où tous les U_i sont iid issues de $\mathcal{U}([0, 1])$.

Le modèle de graphe aléatoire associé à la fonction graphon est généralement appelé modèle de W-graphe dans la littérature et nous utilisons cette appellation dans ce mémoire. Dans certains travaux, ce modèle est directement désigné par modèle graphon. Construit uniquement à partir de la notion d'échangeabilité, il généralise pratiquement tous les modèles existants de graphe aléatoire, en particulier tous les dérivés (binaires) de SBM. Ces dix dernières années, une part importante de la recherche en analyse statistique des réseaux s'est donc tournée vers le W-graphe et beaucoup d'efforts ont été fait afin de proposer des techniques d'inférence. Notons que ce modèle peut également être introduit comme la limite d'une séquence de réseaux denses, où le nombre de connexions croit quadratiquement avec le nombre de nœuds. En effet, Lászlo et Szegedy, 2006 ont montré que toutes séquences de réseaux denses a pour caractérisation limite un réseau suivant un W-graphe. De ce résultat naissent les critiques récentes de ce modèle : les réseaux issus d'un W-graphe sont denses ou vides. Il s'agit essentiellement d'une conséquence de l'application de la loi forte des grands nombres pour les U statistiques (Arcones et Gine, 1992 ; Gine et Zinn, 1992 ; Hoeffding, 1961). Or, beaucoup d'études pratiques considèrent les réseaux comme creux, c'est-à-dire que peu de connexions sont réellement présentes, au regard du nombre possible. Pour plus de détails, nous renvoyons le lecteur à Caron et Fox, à paraître.

4.2 Estimation dans le modèle W-graphe

Nous présentons ici les travaux publiés dans Latouche et Robin, 2016 permettant de réaliser l'estimation de la fonction graphon d'un W-graphe et des probabilités de motifs. Le point de départ de la stratégie d'inférence est le même que celui utilisé par les autres approches proposées dans la littérature. SBM étant un cas particulier de W-graphe pour lequel des techniques d'estimation existent, l'idée est de s'appuyer sur ce modèle comme proxy.

Soient $\pi = (\pi_q)_q$ un vecteur de probabilités telles que $\sum_{q=1}^{Q} \pi_q = 1$ et $\mu = (\mu_{qr})_{qr}$ une matrice symétrique de probabilités de taille $Q \times Q$. En notant,

$$\sigma_q = \sum_{j=1}^{q} \pi_j,$$

ainsi que

$$C_\pi(u) = 1 + \sum_{q=1}^{Q} \mathbb{1}_{[\sigma_q, +\infty[}(u),$$

et en définissant

$$W(u, v) = \mu_{C(u),C(v)},$$

le W-graphe construit est un SBM avec pour paramètres π et μ. L'intervalle $[0, 1]$ est ainsi découpé en Q segments, chaque segment étant associé à un cluster particulier désigné par $C_\pi(\cdot)$. Le graphon est alors constant par morceaux comme illustré par la figure 4.1.

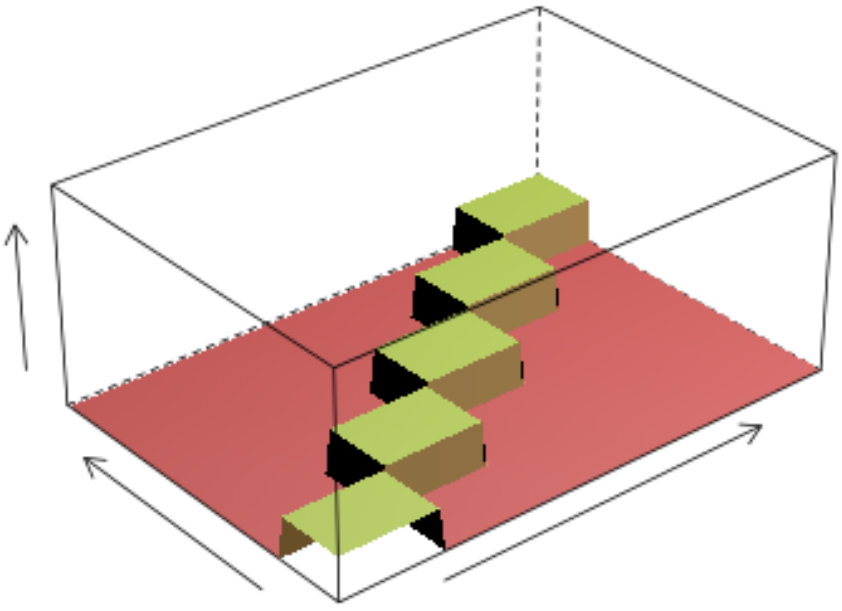


Figure 4.1 – La fonction graphon, constante par morceaux, associée à un SBM.

IDENTIFIABILITÉ Le modèle de W-graphe n'est pas identifiable. En effet, soient une fonction graphon $h_1(\cdot,\cdot)$ ainsi qu'une transformation $\rho(\cdot)$ de $[0,1]$, préservant la mesure, alors le graphon $h_2(\rho(u),\rho(v))$ induit la même mesure sur l'ensemble des graphes à n nœuds que $h_1(u,v)$. Une contrainte simple permet de caractériser la classe d'équivalence associée par un unique représentant. Il suffit de choisir une fonction graphon $W(\cdot,\cdot)$ telle que le degré moyen $D(u) = \int W(u,v)dv$ soit une fonction croissante monotone. C'est le choix que nous avons fait dans toutes les expériences que nous avons réalisées. Dans le cas d'un SBM, cela implique que $d_q = \sum_\ell \pi_\ell \mu_{q\ell}$ augmente avec q.

4.2.1 Estimation de la fonction graphon

À nombre de clusters Q fixe, l'algorithme VBEM introduit en section 2.2 est employé pour estimer la loi a posteriori de π et μ, sachant les données dans X. Cette loi, issue d'une approximation variationnelle, peut ensuite être intégrée afin d'estimer la loi a posteriori de la fonction graphon, en (u,v).

Proposition 4.1. *Pour $(u,v) \in [0,1]^2$, $u \leq v$, et en utilisant un SBM à Q clusters, l'approximation variationnelle $\widetilde{p}(W(u,v)|X,Q)$ de la loi a posteriori de $W(u,v)$, sachant les données dans X, peut être calculée en $\mathcal{O}(Q^2)$.*

Le point clé vient de l'algorithme de récursion de GOUDA et SZÁNTAI, 2010 permettant de calculer efficacement des fonctions de répartition de lois de Dirichlet. Nous donnons dans la proposition suivante l'espérance de cette loi approchée. Les écarts types peuvent également être déterminés.

Proposition 4.2. *Pour $(u,v) \in [0,1]^2$, $u \leq v$, et en utilisant un SBM à Q clusters, l'approximation variationnelle de l'espérance de la loi a posteriori de $W(u,v)$, sachant X, est donnée par :*

$$\widetilde{\mathrm{E}}[W(u,v)|X,Q] = \sum_{q\leq\ell}^{Q} \frac{\eta_{q\ell}}{\eta_{q\ell}+\zeta_{q\ell}} \left[F_{q-1,\ell-1}(u,v;a) - F_{q,\ell-1}(u,v;a) - F_{q-1,\ell}(u,v;a) + F_{q,\ell}(u,v;a)\right],$$

où $F_{q,\ell}(u,v;a)$ désigne la fonction de répartition jointe du couple de variables aléatoires (σ_q,σ_ℓ) lorsque π suit la loi de Dirichlet $\mathrm{Dir}(a)$. Les paramètres dans a correspondent aux n_q introduits en section 2.2. Les matrices $\eta = (\eta_{ql})_{ql}$ et $\eta = (\zeta_{ql})_{ql}$ sont également données dans cette section.

AGRÉGATION DE MODÈLES L'algorithme VBEM de LATOUCHE, BIRMELE et AMBROISE, 2012 permet également d'estimer la log-vraisemblance intégrée des données à l'aide de la borne inférieure variationnelle $\mathcal{L}_Q$, à convergence. En considérant une loi a priori uniforme sur Q, l'approximation variationnelle de la loi a posteriori de Q, sachant X, s'écrit donc :

$$\tilde{p}(Q|X) \propto \exp\left(\mathcal{L}_Q\right),$$

telle que $\sum_{q=1}^{Q} \tilde{p}(Q|X) = 1$. Cette loi peut être intégrée afin d'agréger de manière pondérée les estimateurs associés à des SBM, pour des Q différents. L'estimateur obtenu est alors de type BMA (HOETING, MADIGAN et al., 1999), Bayesian model averaging en anglais.

Proposition 4.3. *Pour $(u,v) \in [0,1]^2$, $u \leq v$, l'estimateur BMA variationnel $\widetilde{p}(W(u,v)|X)$ de la loi a posteriori de $W(u,v)$, sachant les données dans X est donné par :*

$$\tilde{p}(W(u,v)|X) = \sum_{Q \geq 1} \tilde{p}(Q|X)\tilde{p}(W(u,v)|X,Q).$$

En pratique, une séquence finie de SBM est utilisée pour l'agrégation car $\tilde{p}(Q|X)$ est numériquement à zéro, pour des valeurs de Q trop élevées.

4.2.2 Probabilités d'apparition de motifs

Comme rappelé précédemment, le modèle de W-graphe souffre de problème d'identifiabilité. Cependant, Diaconis et Janson, 2008 ont montré que le nombre d'occurrences d'un motif était invariant et par conséquent caractéristique du modèle. Un motif peut être défini comme un sous-graphe avec des arêtes particulières. Plus précisément, le motif $\mathbf{m}$ à k nœuds est représenté par sa matrice d'adjacence telle que $m_{ij} = 1$ si les nœuds i et j sont connectés, 0 sinon. L'occurence de $\mathbf{m}$ en une position $\beta = (i_1, \ldots, i_k)$, avec $i_1 < \cdots < i_k$, est alors donnée par la variable :

$$Y_\beta(\mathbf{m}) = \prod_{1 \leq a < b \leq k} (X_{i_a,i_b})^{m_{ab}}.$$

Ci-dessous $\nu(\mathbf{m})$ désigne la probabilité d'occurence $\mathbb{P}(Y_\beta(\mathbf{m} = 1))$ du motif dans le réseau. L'algorithme VBEM pour SBM est à nouveau utilisé ci-dessous afin d'estimer $\eta(\mathbf{m})$.

Proposition 4.4. *En utilisant un SBM à Q clusters, une approximation variationnelle permet d'estimer l'espérance de la probabilité a posteriori, sachant X, d'apparition d'un motif $\mathbf{m}$:*

$$\begin{aligned}
&\widetilde{\mathrm{E}}[\nu(\mathbf{m})|X,Q] \\
&= \left\{ \left[\prod_{q \leq \ell}^{Q} \frac{\Gamma(\eta_{q\ell} + \zeta_{q\ell})}{\Gamma(\eta_{q\ell})} \right] \frac{\Gamma(\sum_{q=1}^{Q} n_q)}{\prod_{q=1}^{Q} \Gamma(n_q)} \right\} \left\{ \sum_c \left[\prod_{q \leq \ell}^{Q} \frac{\Gamma(\eta_{q\ell} + \eta_{q\ell}^c)}{\Gamma(\eta_{q\ell} + \eta_{q\ell}^c + \zeta_{q\ell})} \right] \frac{\prod_{q=1}^{Q} \Gamma(n_q + n_q^c)}{\Gamma\left[\sum_{q=1}^{Q}(n_q + n_q^c)\right]} \right\},
\end{aligned}$$

où $c = (c_1, \ldots, c_k)$, $n_q^c = \sum_a \mathbb{1}_{\{c_a = q\}}$, $\eta_{q\ell}^c = \sum_{1 \leq a \neq b \leq k} \mathbb{1}_{\{c_a = q\}} \mathbb{1}_{\{c_b = \ell\}} m_{ab}$ pour $q \neq \ell$, $\eta_{qq}^c = \sum_{1 \leq a < b \leq k} \mathbb{1}_{\{c_a = q\}} \mathbb{1}_{\{c_b = q\}} m_{ab}$. Chaque terme c_k prend ses valeurs dans $\{1, \ldots, Q\}$ afin d'indiquer le cluster du nœud k du motif. Toutes les configurations sont testées pour c.

Enfin, comme à la section précédente, un estimateur BMA s'obtient à l'aide d'une agrégation de modèles SBM.

Proposition 4.5. *Un estimateur BMA variationnel de l'espérance de la probabilité a posteriori d'apparition d'un motif $\mathbf{m}$, sachant X, est donné par :*

$$\widetilde{\mathrm{E}}[\nu(\mathbf{m})|X] = \sum_{Q \geq 1} \widetilde{p}(Q|X)\widetilde{\mathrm{E}}[\nu(\mathbf{m})|X,Q]. \tag{4.1}$$

4.2.3 Expérimentations numériques

Nous présentons ici les résultats sur simulations numériques obtenus pour l'évaluation de la qualité de la procédure d'estimation de la fonction graphon. Ainsi, des réseaux sont générés à partir d'un modèle de W-graphe défini par le graphon produit $W(u,v) = g(u)g(v)$ où :

$$g(u) = \sqrt{\rho}\lambda u^{\lambda - 1}.$$

Le paramètre ρ contrôle la densité des réseaux, c'est-à-dire la probabilité moyenne de connexion entre deux nœuds quelconques, alors que λ détermine la concentration des degrés : plus la valeur de λ est élevée, plus les arêtes sont concentrées autour de quelques nœuds. Plusieurs configurations sont testées avec n entre 100 et 316, $\log_{10}\rho \in \{-2, -1.5, -1\}$, et $\lambda \in \{1, 2, 3, 5\}$. Pour chacune d'entre elles, 100 réseaux sont simulés et la procédure d'estimation appliquée. Les résultats sont évalués à l'aide de la racine de l'erreur carrée moyenne :

$$RECM = \sqrt{\iint \left[W(u,v) - \widehat{W}(u,v)\right]^2 dudv},$$

où $\widehat{W}(u,v)$ est donnée par la proposition 4.3. Ils sont illustrés sous la forme de boîtes à moustaches en figure 4.2. Les RECM sont globalement proches de zéro, lorsque la densité ou le nombre de nœuds augmentent. La qualité de l'estimation diminue pour des valeurs de λ élevées.

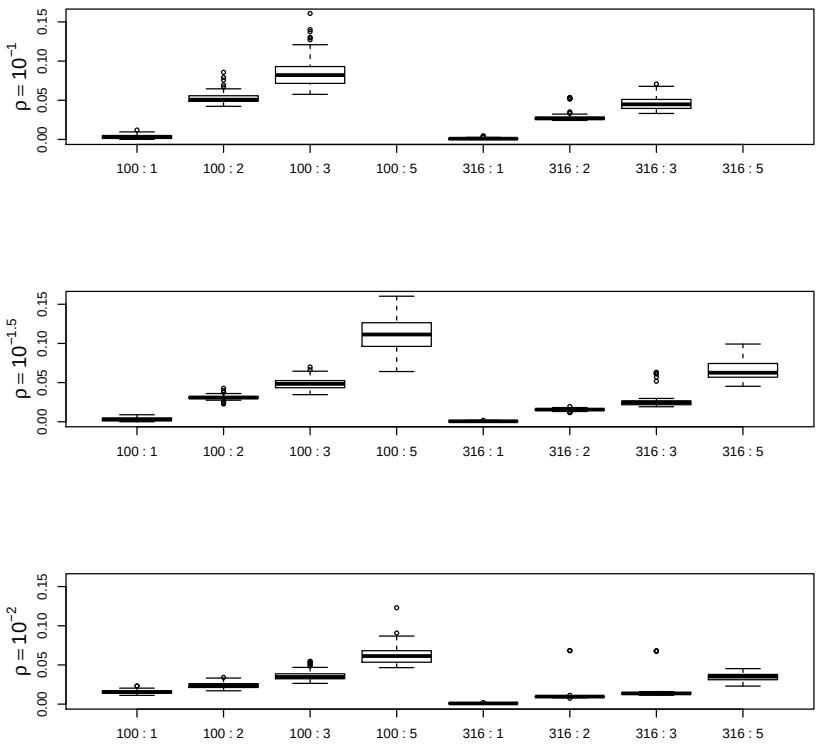


Figure 4.2 – $RECM$ de la fonction graphon estimée pour $\rho = 10^{-2}$, $10^{-1.5}$ et 10^{-1}. Axe des abscisses : nombre de nœuds n et paramètre λ notés $n : \lambda$.

4.3 Qualité de l'ajustement : approche Bayésienne

La régression logistique est une méthode simple permettant de caractériser un réseau binaire à partir de covariables. Une fois les paramètres estimés, le praticien cherche alors à évaluer la qualité de l'ajustement de la régression, afin de tester si les covariables dont il dispose expliquent entièrement la topologie du réseau, et, si ce n'est pas le cas, d'analyser la structure résiduelle. Dans Latouche, Robin et Ouadah, à paraître, nous avons proposé un cadre méthodologique apportant une solution à ce problème. Le modèle de W-graphe généralisant la plupart des modèles existants de graphe aléatoire, l'idée fondatrice de ce travail est d'insérer une fonction graphon dans le modèle standard de régression logistique, afin de caractériser le terme résiduel.

4.3.1 Régression logistique et structure résiduelle

Le modèle de régression logistique permet d'expliquer les connexions binaires dans $X = (X_{ij})_{ij}$ à l'aide de covariables dans $Y = (y_{ij})_{ij}$ où $y_{ij} \in \mathbb{R}^d$:

$$H_0 : \qquad X_{ij}|y_{ij}, \beta, \alpha \sim \mathcal{B}\left[g(y_{ij}^\intercal \beta + \alpha)\right],$$

avec $\beta \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$, et g désigne la fonction logistique $g(t) = 1/(1/\exp(-t))), \forall t \in \mathbb{R}$. Rappelons que, connaissant les covariables dans Y, les connexions sont supposées indépendantes. L'objectif est ici de tester si le modèle H_0 est suffisant pour expliquer la topologie observée du réseau. Pour évaluer la qualité de l'ajustement de H_0, un modèle générique alternatif est introduit :

$$H_1 : \qquad X_{ij}|y_{ij}, \beta, \phi, U_i, U_j \sim \mathcal{B}\left[g(y_{ij}^\intercal \beta + \phi(U_i, U_j))\right],$$

où les variables U_i sont iid tirées à partir de la loi uniforme continue $\mathcal{U}([0,1])$. En retirant le terme de régression $y_{ij}^\intercal \beta$, notons que H_1 correspond exactement au modèle de W-graphe ayant pour fonction graphon $g \circ \phi$. La fonction $\phi(\cdot,\cdot)$ modélise alors la structure résiduelle n'ont expliquée par H_0. Le modèle H_0 est un cas particulier de H_1 où la fonction est constante.

Réaliser l'inférence de H_1 directement est un problème dur à cause du terme de W-graphe. À l'image des travaux présentés dans la section précédente, H_1 est donc remplacé par une série de modèles construits à partir de termes de type SBM avec $Q \geq 1$ blocs :

$$M_Q : \qquad X_{ij}|y_{ij}, \beta, \alpha, Z_i, Z_j \sim \mathcal{B}\left[g(y_{ij}^\intercal \beta + Z_i^\intercal \alpha Z_j)\right], \tag{4.2}$$

où les vecteurs binaires Z_i sont iid supposés issus d'une loi multinomiale de paramètres 1 et $\pi = (\pi_q)_q$. Comme pour un SBM standard, π_q est la probabilité pour un nœud d'être dans le cluster q. De plus, la matrice $\alpha = (\alpha_{qr})_{qr}$ est réelle de taille $Q \times Q$. Nous notons alors $\theta = (\beta, \pi, \alpha)$ l'ensemble des paramètres de M_Q. À nouveau, en l'absence de covariables, M_Q est exactement un modèle SBM de matrice de connectivité μ telle que $\mu_{qr} = g(\alpha_{qr})$.

Nous observons finalement que H_0 est équivalent à M_1 donc le problème d'ajustement se réécrit comme la comparaison entre :

$$H_0 = M_1 \qquad \text{et} \qquad H_1' = \bigcup_{Q \geq 2} M_Q.$$

4.3.2 Comparaison Bayésienne de modèles

Pour la comparaison des modèles, un cadre Bayésien est retenu. Chaque modèle M_Q est donc associé à une loi a priori $p(M_Q)$ et θ est caractérisé par une loi a priori $p(\theta|M_Q)$,

conditionnellement à M_Q. Ensuite, à θ fixe et à covariables dans Y connues, M_Q est supposé avoir généré les connexions dans X. Le cadre Bayésien autorise alors la comparaison des modèles à l'aide des lois a posteriori :

$$p(M_Q|X) = \frac{p(X|M_Q)p(M_Q)}{p(X)} = \frac{p(X|M_Q)p(M_Q)}{\sum_{Q' \geq 1} p(X|M_{Q'})p(M_{Q'})}. \tag{4.3}$$

Notons que la dépendance en Y n'est pas indiquée dans ces expressions, car évidente. La qualité de l'ajustement de H_0 peut ainsi s'évaluer grâce à $p(H_0|X) = p(M_1|X)$. Le facteur de Bayes (Kass et Raftery, 1995) peut également être calculé :

$$B_{01} = \frac{p(X|H_0)}{p(X|H_1')} \qquad \text{où} \quad p(X|H_1') = \frac{1}{p(H_1')} \sum_{Q \geq 2} p(M_Q)p(X|M_Q).$$

Lois a priori Sans connaissance particulière a priori sur les modèles, nous donnons le même poids $p(H_0) = p(H_1') = 1/2$ pour H_0 et H_1'. Par conséquent, $p(M_1) = 1/2$. Pour le modèle M_Q, un produit de lois a priori conjuguées est utilisé pour modéliser θ : $p(\theta|M_Q) = p(\beta|M_Q)p(\pi|M_Q)p(\alpha|M_Q)$. Le vecteur π intervenant dans une loi multinomiale, nous considérons une loi de Dirichlet :

$$p(\pi|M_Q) = \mathrm{Dir}(\pi; e).$$

Le choix des hyperparamètres dans le vecteur $e = (e_1, \ldots, e_Q)$ est discuté en section 2.2 de ce mémoire. Ensuite, le vecteur β est caractérisé par une loi a priori normale :

$$p(\beta|\eta, M_Q) = \mathcal{N}\left(\beta;\, 0, \frac{I_d}{\eta}\right) = \prod_{j=1}^{d} \mathcal{N}\left(\beta_j;\, 0, \frac{1}{\eta}\right),$$

où $\eta > 0$ est un paramètre contrôlant l'inverse variance. De manière similaire, un produit de lois a priori normale d'inverse variance $\gamma > 0$ est employé pour α :

$$p(\alpha|\gamma, M_Q) = \prod_{k \leq l}^{Q} \mathcal{N}\left(\alpha_{kl};\, 0, \frac{1}{\gamma}\right).$$

Comme des réseaux non orientés sont ici considérés, une contrainte de symétrie est imposée à α. Finalement, les paramètres η et γ sont modélisés par des lois Gamma :

$$p(\gamma|M_Q) = \mathrm{Gam}(\gamma;\, a_0, b_0), \quad a_0, b_0 > 0,$$

et

$$p(\eta|M_Q) = \mathrm{Gam}(\eta;\, c_0, d_0), \quad c_0, d_0 > 0.$$

Pour le choix des hyperparamètres a_0, b_0, c_0, et d_0, nous renvoyons le lecteur à la section 2.3.3 Les lois $p(\beta|Q)$ et $p(\alpha|Q)$ appartiennent alors à la famille des distributions hyperboliques généralisées. Cette famille est discutée en profondeur dans Caron et Doucet, 2008.

4.3.3 Inférence

Les termes $p(M_Q|X)$ dans (4.3) font intervenir les vraisemblances intégrées des données $p(X|Q)$ qui n'ont malheureusement pas de forme analytique. Dans Latouche, Robin et Ouadah, à paraître, nous avons donc introduit deux approximations variationnelles ainsi

qu'un algorithme d'optimisation permettant de construire une borne inférieure de $\log p(X|Q)$. L'algorithme alterne entre des étapes d'optimisation fonctionnelle par rapport à une loi r, estimation de $p(Z, \pi, \alpha, \beta, \gamma, \eta|X, Q)$, et une matrice $\xi = (\xi_{ij})_{ij}$ de paramètres utilisés pour des approximations locales de la fonction logistique.

Proposition 4.6. *Pour une loi* $r(Z, \pi, \alpha, \beta, \gamma, \eta) = r(\pi)r(\alpha)r(\beta)r(\gamma)r(\eta)\prod_{i=1}^{n} r(Z_i)$ *et une matrice* $\xi = (\xi_{ij})_{ij}$ *positive réelle de taille* $n \times n$*, obtenues à convergence de l'algorithme d'optimisation de* Latouche, Robin *et* Ouadah, *à paraître, une borne inférieure de la log-vraisemblance marginale est donnée par :*

$$\begin{aligned}\mathcal{L}_Q(r;\,\xi) = \frac{1}{2}\sum_{i\neq j}^{n}\left\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij})\xi_{ij}^2\right\} + \log\frac{C(e^n)}{C(e)} + \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + \log\frac{\Gamma(c_n)}{\Gamma(c_0)}\\ + a_0\log b_0 + a_n(1 - \frac{b_0}{b_n} - \log b_n) + c_0\log d_0 + c_n(1 - \frac{d_0}{d_n} - \log d_n)\\ + \frac{1}{2}\sum_{q\leq l}^{Q}\log(\sigma_\alpha^2)_{ql} + \frac{1}{2}\log|S_\beta| - \sum_{i=1}^{n}\sum_{q=1}^{Q}\tau_{iq}\log\tau_{iq} + \frac{1}{2}\sum_{q\leq l}^{Q}(\sigma_\alpha^2)_{ql}^{-1}(m_\alpha)_{ql}^2 - \frac{1}{2}m_\beta^\intercal S_\beta^{-1}m_\beta\\ + \frac{1}{2}m_\beta^\intercal\sum_{i\neq j}^{n}(X_{ij} - \frac{1}{2})y_{ij},\end{aligned}$$

où $\lambda(x) = (g(x) - 1/(2x))/(2x)$ *et* $C(x) = \prod_{q=1}^{Q}\Gamma(x_q)\Big/\Gamma\left(\sum_{q=1}^{Q}x_q\right)$ *avec* $\xi_{ij} \in \mathbb{R}^+, \xi_{ij} = \xi_{ji}$*. De plus*

$\tau_i = (\tau_{i1}, \ldots, \tau_{iQ})^\intercal$ **:**

$$\begin{aligned}\tau_{iq} \propto \exp\left\{\sum_{l=1}^{Q}(m_\alpha)_{ql}\sum_{j\neq i}^{n}\Big((X_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij})y_{ij}^\intercal m_\beta\Big)\tau_{jl} - \sum_{l=1}^{Q}\mathrm{E}_{\alpha_{ql}}[\alpha_{ql}^2]\sum_{j\neq i}^{n}\lambda(\xi_{ij})\tau_{jl}\right.\\ \left. + \psi(e_q^n) - \psi\Big(\sum_{l=1}^{Q}e_l^n\Big)\right\},\end{aligned}$$

tel que $\sum_{q=1}^{Q}\tau_{iq} = 1$.

$e^n = (e_1^n, \ldots, e_Q^n)$ **:**

$$e_q^n = e_0 + \sum_{i=1}^{n}\tau_{iq}.$$

S_β^{-1} **:**

$$S_\beta^{-1} = \frac{c_n}{d_n}I_d + \sum_{i\neq j}^{n}\lambda(\xi_{ij})y_{ij}y_{ij}^\intercal.$$

m_β **:**

$$m_\beta = S_\beta\frac{1}{2}\sum_{i\neq j}^{n}\left(X_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij})\tau_i^\intercal m_\alpha\tau_j\right)y_{ij}.$$

a_n **:**

$$a_n = a_0 + \frac{Q(Q+1)}{4}.$$

b_n :

$$b_n = b_0 + \frac{1}{2}\sum_{q\leq l}^{Q} \mathrm{E}_{\alpha_{ql}}[\alpha_{ql}^2].$$

c_n :

$$c_n = c_0 + \frac{d}{2}.$$

d_n :

$$d_n = d_0 + \frac{1}{2}\mathrm{Tr}(S_\beta) + \frac{1}{2}m_\beta^\intercal m_\beta.$$

$(\sigma_\alpha)_{ql}^{-1}$:

$$(\sigma_\alpha^2)_{qq}^{-1} = \frac{a_n}{b_n} + \sum_{i\neq j}^{n} \lambda(\xi_{ij})\tau_{iq}\tau_{jq}, \forall q,$$

$$(\sigma_\alpha^2)_{ql}^{-1} = \frac{a_n}{b_n} + 2\sum_{i\neq j}^{n} \lambda(\xi_{ij})\tau_{iq}\tau_{jl}, \forall q \neq l,$$

$(m_\alpha)_{ql}$:

$$(m_\alpha)_{qq} = (\sigma_\alpha^2)_{qq}\sum_{i\neq j}^{n}\left(\frac{1}{2}(X_{ij} - \frac{1}{2}) - \lambda(\xi_{ij})y_{ij}^\intercal m_\beta\right)\tau_{iq}\tau_{jq}, \forall q,$$

$$(m_\alpha)_{ql} = (\sigma_\alpha^2)_{ql}\sum_{i\neq j}^{n}\left((X_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij})y_{ij}^\intercal m_\beta\right)\tau_{iq}\tau_{jl}, \forall q \neq l.$$

Le développement de l'algorithme d'optimisation et de la borne font intervenir des éléments techniques liés à ceux rencontrés pour le modèle OSBM. Comme décrit en section 2.3, les difficultés apparaissent à cause de la fonction logistique qui est non linéaire et du modèle graphique. L'algorithme est donné dans Latouche, Robin et Ouadah, à paraître. À convergence, une approximation variationnelle de $p(M_Q|X)$ s'obtient alors par :

$$\widehat{p}(M_Q|X) \propto p(M_Q)\exp\{\widehat{\mathcal{L}}_Q\},$$

de manière à ce que $\sum_{q=1}^{Q}\widehat{p}(M_Q|X) = 1$.

Résidu L'objectif premier est ici d'estimer $p(M_Q|X)$ afin d'évaluer la qualité d'estimation d'un modèle de régression logistique. L'algorithme décrit précédemment permet également d'obtenir naturellement un estimateur du terme de type W-graphe dans H_1.

Proposition 4.7. *Pour $(u,v) \in [0,1]^2, u \leq v$, un estimateur BMA variationnelle de l'espérance de la loi a posteriori du résidu ϕ, sachant X, est donné par :*

$$\widehat{\mathrm{E}}\big[\phi(u,v)|X\big] = \sum_{Q\geq 1}\widehat{p}(M_Q|X)\widehat{\mathrm{E}}\big[\phi(u,v)|X, M_Q\big],$$

où

$$\widehat{\mathrm{E}}\big[\phi(u,v)|Y, M_Q\big] = \sum_{q\leq l}^{Q}(m_\alpha)_{ql}\left[F_{q-1,l-1}(u,v;e^n) - F_{q,l-1}(u,v;e^n) - F_{q-1,l}(u,v;e^n) + F_{q,l}(u,v;e^n)\right].$$

Ce résultat est essentiellement basé sur nos travaux dans Latouche et Robin, 2016 et décrit à la section 4.2. À nouveau, $F_{q,\ell}(u, v; e^n)$ désigne la fonction de répartition jointe du couple de variables (σ_q, σ_l), telles que $\sigma_q = \sum_{l=1}^{q} \pi_l$, et π suit une loi de Dirichlet $\mathrm{Dir}(e^n)$.

4.3.4 Expérimentations numériques

Les résultats d'évaluation sur simulations numériques de l'estimation de $p(H_0|X) = p(M_1|X)$ sont ici présentés. Tous les réseaux sont tirés à partir du modèle H_1. Ainsi, chaque nœud est associé à une position latente U_i générée à partir de la loi $\mathcal{U}([0, 1])$, uniforme continue sur $[0, 1]$. Ensuite, un vecteur $y_i \in \mathbb{R}^d$ de covariables est simulé pour chaque nœud, à partir d'une loi normale multivariée centrée réduite avec $d = 2$. Chaque vecteur y_{ij}, pour $i < j$, est alors fixé à $y_{ij} = y_i - y_j$. La fonction ϕ utilisée s'écrit $g^{-1}(W(u, v))$ où $W(u, v) = \rho\lambda^2(uv)^{\lambda-1}$ comme à la section 4.2. Notons que $\lambda = 1$ correspond à un résidu constant, c'est-à-dire à H_0. En revanche, lorsque $\lambda > 1$, le modèle de simulation est H_1. Plusieurs configurations sont testées avec $n \in \{100, 150\}, \rho \in \{10^{-2}, 10^{-1.5}, 10^{-1}\}$, et λ prenant ses valeurs dans $[1, 5]$. Pour chaque configuration, 100 réseaux sont générés et l'estimateur $\widehat{p}(H_0|X)$ est construit. Les résultats sont illustrés en figure 4.3. Les seuils de détection de H_1 sont bas et la discrimination entre H_0 et H_1 devient plus nette à mesure que n et ρ augmentent.

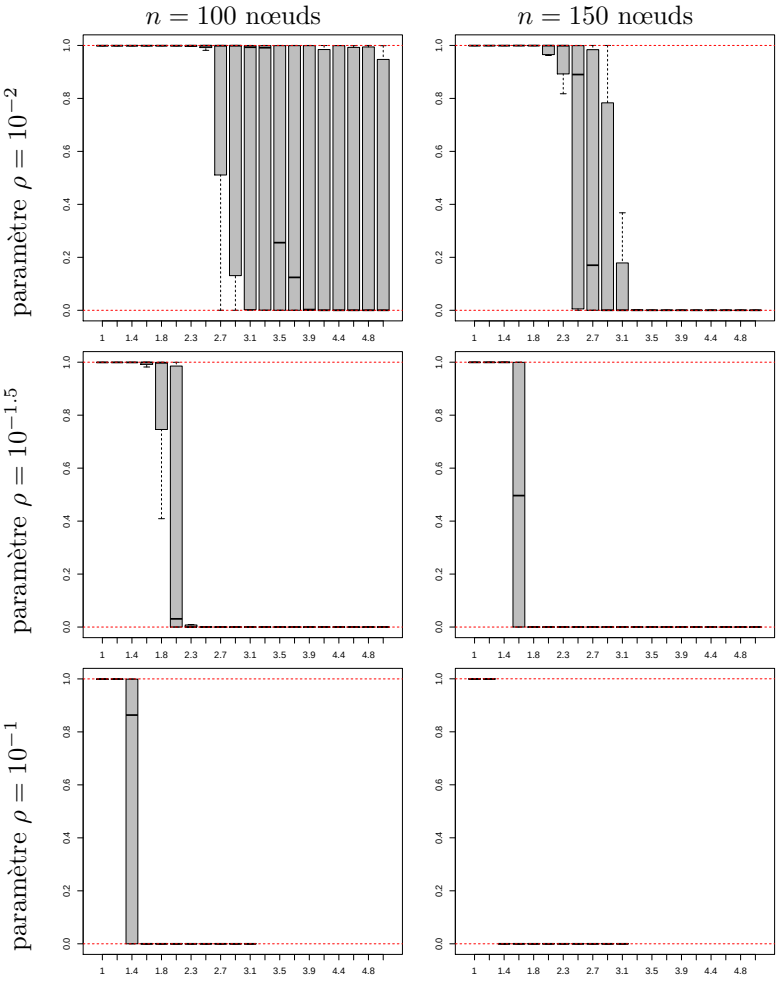


Figure 4.3 – Boîtes à moustaches des valeurs de l'estimateur $\hat{p}(H_0|Y)$, pour des valeurs de λ entre 1 et 5. Six scénarios sont considérés avec $n \in \{100, 150\}$ et $\rho \in \{10^{-2}, 10^{-1.5}, 10^{-1}\}$. Le modèle H_0 est vrai pour $\lambda = 1$ et faux pour $\lambda > 1$.

4.4 Qualité de l'ajustement : tests

Nous présentons ici nos travaux introduits dans Ouadah, Robin et Latouche, 2017. À l'image de la section précédente, l'objectif est d'évaluer la qualité d'ajustement de modèles de graphe aléatoire. Deux modèles particuliers sont considérés : le modèle indépendant qui inclut le modèle standard d'Erdös et Rényi, 1959 ainsi que le modèle de W-graphe qui est l'objet de ce chapitre et qui généralise la plupart des modèles existants, comme discuté précédemment. Des tests sont construits à partir d'une statistique basée sur les degrés observés, c'est-à-dire sur le nombre de connexions de chaque nœud :

$$T_{\theta_0} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \mathrm{E}_{\theta_0}[D_i])^2. \tag{4.4}$$

D_i désigne le degré du nœud i et $\mathrm{E}_{\theta_0}[D_i]$ est l'espérance de D_i, sous un modèle de graphe aléatoire de paramètre θ_0. Pour les deux types de modèles, nous avons en autre prouvé la normalité asymptotique de la statistique T_{θ_0}.

4.4.1 Modèle indépendant

Dans ce modèle, appelé ci-dessous modèle hétérogène d'Erdös-Rényi (HER), toutes les arêtes sont tirées de manière indépendante à l'aide de probabilités fonctions des paires de nœuds considérées. Ainsi, l'arête X_{ij} est supposée générée à partir d'une loi de Bernoulli de probabilité p_{ij}. Par la suite, $p = (p_{ij})_{ij}$ désigne la matrice des p_{ij} et le modèle associé est noté $HER(p)$. Notons que le modèle d'Erdös-Rényi correspond à $p_{ij} = \mu, \forall (i,j)$. Dans ce cadre, la statistique (4.4) s'écrit alors :

$$T_{p^0} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \mu_i^0)^2,$$

avec $D_i = \sum_{j\neq i}^{n} X_{ij}$ et $\mu_i^0 = \sum_{j\neq i}^{n} p_{ij}^0$ est l'espérance de D_i, sous $HER(p^0)$.

Théorème 4.1. *Sous le modèle $HER(p)$, la statistique T_{p^0} est asymptotiquement normale :*

$$(T_{p^0} - \mathrm{E}_p[T_{p^0}])/\mathrm{S}_p[T_{p^0}] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où S_p désigne l'écart type sous $HER(p)$ et

$$\mathrm{E}_p[T_{p^0}] = \frac{2}{n}\left(\sum_{1\leq i<j\leq n}(\sigma_{ij}^2 + \delta_{ij}^2) + \sum_{1\leq i<j<k\leq n}(\delta_{ij}\delta_{ik} + \delta_{ij}\delta_{jk} + \delta_{ik}\delta_{jk})\right), \tag{4.5}$$

avec $\sigma_{ij}^2 = p_{ij}(1-p_{ij})$ et $\delta_{ij} = p_{ij} - p_{ij}^0$. De plus

$$\begin{aligned}\mathrm{V}_p[T_{p^0}] &= \mathrm{S}_p^2[T_{p^0}] \\ &= \frac{4}{n^2}\Big(\sum_{1\leq i<j\leq n}\sigma_{ij}^2(1 - 2p_{ij} + \Delta_i + \Delta_j)^2 + \sum_{1\leq i<j<k\leq n}\left(\sigma_{ij}^2\sigma_{ik}^2 + \sigma_{ij}^2\sigma_{jk}^2 + \sigma_{ik}^2\sigma_{jk}^2\right)\Big),\end{aligned}$$

avec $\Delta_i = \sum_{j\neq i}\delta_{ij}$.

La preuve du théorème 4.1 s'appuie essentiellement sur la décomposition de Hoeffding (voir Van der Vaart, 2000, par exemple) de la statistique T_{p^0} et de l'application du théorème de Lindeberg-Lévy (Billingsley, 1968). Notons que dans Ouadah, Robin et Latouche, 2017, nous avons également donné un résultat similaire pour une autre statistique de test V basée sur les degrés :

$$V = \sum_{i=1}^{n} \frac{1}{n} \left(D_i - \bar{D}\right)^2,$$

où $\bar{D} = (1/n) \sum_{i=1}^{n} D_i$.

Test et puissance Un test est alors mis en place pour comparer deux modèles $H_0 = HER(p^0)$ et $H_1 = HER(p)$ à partir de T_{p^0}. Ainsi, H_0 se voit rejetée dès que la statistique de test dépasse le seuil $\mathrm{E}_{p^0}[T_{p^0}] + t_\alpha \mathrm{S}_{p^0}[T_{p^0}]$ où t_α désigne le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. La puissance du test est donnée par le corollaire suivant.

Corollaire 4.1.1. *La puissance asymptotique du test H_0 contre H_1 est :*

$$\pi(p) = 1 - \Phi\left(\left(\mathrm{E}_{p^0}[T_{p^0}] + t_\alpha \mathrm{S}_{p^0}[T_{p^0}] - \mathrm{E}_p[T_{p^0}]\right) / \mathrm{S}_p[T_{p^0}]\right),$$

où $\Phi(\cdot)$ est la fonction de répartition empirique de la loi normale centrée réduite.

Le corollaire suivant apporte alors une condition suffisante où $HER(p)$ s'écarte de $HER(p^0)$ asymptotiquement, permettant ainsi au test de les discerner.

Corollaire 4.1.2. *Considérons deux matrices p^0, p, et définissons :*

$$\Delta_n(p^0, p) = \frac{2}{n} \left(\sum_{1 \le i < j \le n} \delta_{ij}^2 + \sum_{1 \le i < j < k \le n} (\delta_{ij}\delta_{ik} + \delta_{ij}\delta_{jk} + \delta_{ik}\delta_{jk}) \right).$$

Si $\Delta_n(p^0, p) = \Theta(n^\alpha)$ et $\alpha > 1/2$, alors le test H_0 contre H_1 est asymptotiquement puissant.

Cas des réseaux creux La validé du théorème 4.1 est maintenant discutée dans le cas de réseaux creux où le nombre d'arêtes présentes est faible par rapport au nombre d'arêtes possibles. Plus précisément, la parcimonie est étudiée de deux manières faisant dépendre les p_{ij} de n. Soit les probabilités de connexion tendent vers 0 à mesure que n augmente, soit la fraction de probabilités différentes de 0 décroit quand n augmente.

Proposition 4.8. *Considérons le modèle $HER(p)$ où $p_{ij} = p_{ij}^* n^{-a}$, $a > 0$, $p_{ij} \in [0,1]$ et une fraction $1 - n^{-b}$, $b \ge 0$, des p_{ij} est mise à 0. Les probabilités p_{ij}^0 sont supposées satisfaire les mêmes conditions. Si $a + b < 2$, alors la statistique est bien asymptotiquement normale.*

Cette preuve s'appuie à nouveau sur les projections obtenues par la décomposition de Hoeffding de T_{p^0} et des calculs d'ordre en n.

4.4.2 Modèle de W-graphe

Nous nous intéressons dans cette section au modèle de W-graphe caractérisé par une fonction graphon $W : [0,1]^2 \to [0,1]$. Pour rappel, ce modèle est construit à partir de la notion

d'échangeabilité. Contrairement au modèle précédent, il ne fait pas l'hypothèse d'indépendance entre les arêtes. Dans ce cadre, la statistique (4.4) s'écrit :

$$T_{W^0} = \frac{1}{n}\sum_i (D_i - (n-1)W_1^0)^2,$$

où $W_1^0 = \int\int W^0(u,v)dudv$ est la probabilité marginale d'apparition d'une arête sous un modèle, noté $GE(W^0)$ (pour graphe échangeable), caractérisé par le graphon $W^0(\cdot,\cdot)$.

Théorème 4.2. *Sous le modèle $GE(W)$, la statistique de test T_{W^0} est asymptotiquement normale :*

$$(T_{W^0} - \mathrm{E}_W[T_{W^0}])/\mathrm{S}_W[T_{W^0}] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

avec

$$\mathrm{E}_W[T_{W^0}] = n^{-1}\left\{n(n-1)^2(W_1^0)^2 + [1-2(n-1)W_1^0]n_1W_1 + n_2W_2\right\}$$

et

$$\begin{aligned}
\mathrm{V}_W[T_{W^0}] &= \mathrm{S}_W^2[T_{W^0}] \\
&= n^{-2}\left\{4[1-2(n-1)W_1^0]^2\left(\frac{n_1}{2}W_1 + n_2W_2 + \frac{n_3}{4}W_1^2 - \frac{n_1^2}{4}W_1^2\right)\right. \\
&\quad + 8[1-2(n-1)W_1^0]\left[\frac{n_2}{2}(2W_2+W_3) + \frac{n_3}{2}(W_5+2W_6) + \frac{n_4}{2}W_1W_2 - \frac{n_1n_2}{4}W_1W_2\right] \\
&\quad + 4\left[\frac{n_2}{6}(3W_2+6W_3) + \frac{n_3}{2}(4W_4+2W_5+2W_6+W_7)\right. \\
&\quad \left.\left. + \frac{n_4}{4}(4W_8+W_9+4W_{10}) + \left(\frac{n_5}{5} - \frac{n_2^2}{4}\right)W_2^2\right]\right\}.
\end{aligned}$$

On a $n_i = \prod_{k=0}^{i}(n-k)$. De plus, W_i désigne la probabilité d'apparition du motif R_i (voir figure 4.4) dans le graphe, sous $EG(W)$.

La preuve s'appuie sur un résultat de convergence (Bickel, Chen et Levina, 2011) des probabilités empiriques de motifs dans le cas du modèle de W-graphe. Pour certains modèles, en particulier SBM, une forme explicite s'obtient pour le calcul des W_i. Pour plus de détails, nous renvoyons le lecteur à Ouadah, Robin et Latouche, 2017.

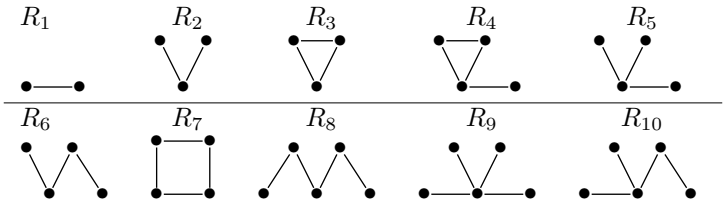


Figure 4.4 – Illustration des motifs R_1 à R_{10} apparaissant dans le calcul des moments de la statistique T_{W^0}.

Test et puissance Le théorème 4.2 permet de mettre en place un test au niveau α asymptotique de $H_0 = EG(W^0)$ contre $H_1 = EG(W)$. Ainsi, H_0 est rejetée dès que la statistique T_{W^0} dépasse $\mathrm{E}_{W^0}[T_{W^0}] + t_\alpha \mathrm{S}_{W^0}[T_{W^0}]$. La puissance associée est donnée par le corollaire suivante.

Corollaire 4.2.1. *La puissance asymptotique du test H_0 contre H_1 est :*

$$\pi(p) = 1 - \Phi\left(\left(\mathrm{E}_{W^0}[T_{W^0}] + t_\alpha \mathrm{S}_{W^0}[T_{W^0}] - \mathrm{E}_W[T_{W^0}]\right)/\mathrm{S}_W T_{W^0}\right).$$

Comme pour le modèle HER, une condition suffisante est considérée pour s'assurer que $EG(W^0)$ s'écarte de $EG(W)$ asymptotiquement afin de les distinguer.

Corollaire 4.2.2. *Soient deux fonctions graphons $W^0(\cdot,\cdot)$ et $W(\cdot,\cdot)$ ainsi que :*

$$\Delta_n(W^0, W) = [1 - 2(n-1)W_1^0]n_1(W_1 - W_1^0) + n_2(W_2 - W_2^0).$$

Si $\Delta_n(W^0, W) > 0$, alors le test $H_0 = EG(W^0)$ contre $H_1 = EG(W)$ est asymptotiquement puissant.

CAS DES RÉSEAUX CREUX La validité du théorème 4.2 est étudiée dans le cas de réseaux creux, où la densité W_1 du réseau varie avec n. Ainsi, $W_1 = W_1(n)$ tend vers 0 à mesure que n augmente, à une vitesse discutée ci-dessous.

Proposition 4.9. *Sous le modèle de W-graphe définit par le graphon $W(\cdot,\cdot)$, tel que W_1 et W_1^0 sont d'ordre $n^{-2/3}$ ou plus, si $\int\int \left(W(u,v)/W_1(n)\right)^6 < \infty$, alors la statistique de test T_{W^0} est bien asymptotiquement normale.*

La preuve est construite à partir de la décomposition utilisée dans le théorème 4.2 de T_{W^0}, du théorème de convergence des motifs de BICKEL, CHEN et LEVINA, 2011, ainsi que du lemme de Slutsky.

4.4.3 Expérimentations numériques

Les travaux de cette section étant essentiellement théoriques, des expérimentions sur données simulées ont été utilisées afin d'illustrer la normalité asymptotique des statistiques de test. Seuls les résultats pour le modèle de W-graphe sont présentés ci-dessous. Ainsi, un graphon $W^0(\cdot,\cdot)$ constant par morceau, caractérisant donc un SBM (voir section 4.2), est utilisé. Plus précisément, une forme produit est retenue, telle que $W^0(u,v) = \eta_q\eta_\ell$ si u est dans le bloc q et v dans ℓ. Deux blocs sont employés et η_1 ainsi que η_2 sont fixés à 0.4 et 0.5, respectivement. Le graphon est alors contaminé par un terme de parcimonie $W(u,v) = W^0(u,v)n^{-a}$. Notons que la densité initiale associée à W^0 est fixée à 0.1. Plusieurs valeurs de (a,n) sont alors utilisées et, pour chaque configuration, 1000 simulations sont faites. La statistique T_{W^0} est alors calculée à chaque fois. Le diagramme quantile-quantile de la figure 4.5 illustre le comportement de T_{W^0} après centrage et réduction. La normalité apparaît ici clairement. En revanche la statistique s'écarte de la loi normale pour des valeurs de a trop importantes.

Conclusion

Ce chapitre s'est intéressé principalement au modèle de W-graphe qui est caractérisé par une fonction appelée graphon. Ce modèle est construit uniquement à partir de la notion d'échangeabilité et généralise la plupart des modèles de graphe aléatoire, comme tous les dérivés simples du modèle SBM. La flexibilité du W-graphe pose de nombreux problèmes pour l'estimation sur données réelles. À partir du constat simple qu'un SBM peut être caractérisé par une fonction graphon constante par morceaux, nous avons dérivé une procédure d'inférence. Elle agrège des modèles SBM avec des niveaux de complexité croissants. Ensuite, le W-graphe a été considéré pour répondre à une question soulevée régulièrement en

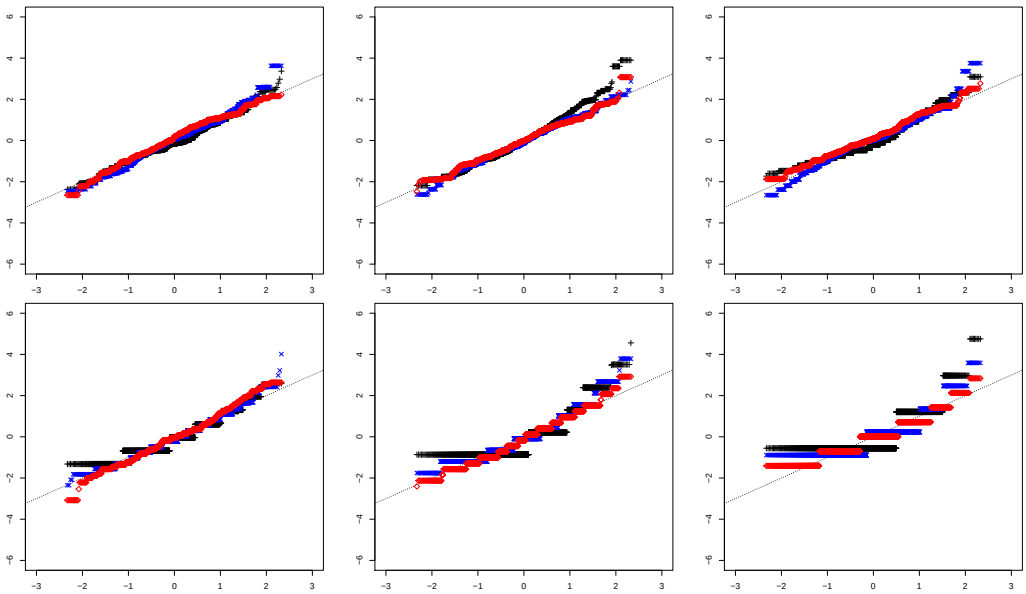


Figure 4.5 – Diagramme quantile-quantile de la statistique de test T_{W^0} sous un modèle de W-graphe, pour des probabilités de connexion décroissantes avec n : $W(u,v) = n^{-a}W^0(u,v)$ et de densité initiale $\rho^* = 0.1$. De en haut à gauche à en bas à droite : $a = 0, 0.4, 0.8, 1.2, 1.4, 1.6$. Valeurs de n= 100 (+), 1 000 ($\times$) et 10 000 ($\diamond$).

analyse des réseaux. Lorsqu'il dispose de covariables, le praticien cherche en effet à savoir si, intégrées dans une régression logistique, ces dernières permettent d'expliquer entièrement la présence de connexions. Le W-graphe est alors utilisé pour modéliser le terme résiduel, non expliqué par la régression. Un cadre d'inférence autorise la mise en place d'un test pour répondre à la question posée. Enfin, des tests ont été mis en place afin de tester l'adéquation entre un modèle de graphe aléatoire et un réseau donné.

5

La grande dimension

5.1 Introduction . . . 68
5.2 Régression linéaire parcimonieuse Bayésienne . . . 68
5.2.1 Le modèle 69
5.2.2 Inférence 70
5.2.3 Sélection de modèles 70
5.2.4 Expérimentations numériques 71
5.3 ACP globalement parcimonieuse . . . 72
5.3.1 Modèle avec bruit 74
5.3.2 Modèle sans bruit 75
5.3.3 Inférence 76
5.3.4 Expérimentations numériques 77
5.4 Sélection de la dimension en ACP Bayésienne . . . 78
5.4.1 Cadre 79
5.4.2 Le choix des hyperparamètres 80
5.4.3 Expérimentations numériques 81

Ce dernier axe de recherche porte sur la modélisation de la grande dimension où les données sont caractérisées par un nombre conséquent de variables. Mes travaux s'organisent en trois sous-axes. Le premier s'intéresse à la régression linéaire parcimonieuse dans un cadre Bayésien. La stratégie proposée, alternative aux procédures de type Lasso, est publiée dans un article de journal (Latouche, Mattei et al., 2016). Le second sous-axe présente les travaux de Bouveyron, Latouche et Mattei, 2017a pour l'estimation de la dimension intrinsèque des données. Enfin, le dernier sous-axe décrit les stratégies de Bouveyron, Latouche et Mattei, 2017b pour l'estimation du nombre optimal d'axes en analyse en composantes principales.

5.1 Introduction

Les outils statistiques développés au cours du 20ième siècle font pour la plupart l'hypothèse que le nombre n d'observations présentes dans les données à analyser est plus important que le nombre p de variables les caractérisant. Or, les dernières décennies ont vu apparaître des données très différentes par nature, suite à l'émergence des nouvelles technologies et en particulier des moyens d'acquisition. Certaines données peuvent ainsi être décrites par des millions de variables. Lorsque p devient conséquent par rapport à n, le cadre statistique est celui de la grande dimension. Des données de ce type apparaissent naturellement dans de nombreux domaines scientifiques, notamment en analyse d'images, dans les biotechnologies, et en chimiométrie. La géométrie des espaces de représentation des observations, lorsque p est grand, est particulièrement complexe. Pour une discussion sur les problèmes géométriques liés à la grande dimension, nous renvoyons le lecteur à Bellman, 1961 sur le fléau de la dimension, et à Scott et Thompson, 1983. Ces derniers rappellent en particulier que le volume de l'hypersphère unité de dimension p tend rapidement vers 0 lorsque p augmente. Par conséquent, les voisinages au sens Euclidien du terme sont tous essentiellement vides. Ce phénomène est appelé phénomène de l'espace vide. Les outils statistiques existants doivent donc être adaptés pour ce type d'espace. Les solutions développées cherchent essentiellement à réduire le nombre de variables ou à projeter les observations dans un espace de dimension plus faible. Notons que la grande dimension apparaît aussi bien en apprentissage statistique supervisé que non supervisé. Les travaux que nous présentons dans ce chapitre proposent donc des éléments de solution pour ces deux cadres d'apprentissage.

5.2 Régression linéaire parcimonieuse Bayésienne

Dans cette section, sont présentés les travaux publiés dans Latouche, Mattei et al., 2016 permettant de faire de la régression linéaire parcimonieuse, dans un cadre de sélection de modèles Bayésienne. Le point de départ est standard. Il s'agit de sélectionner les variables utiles à la prédiction (linéaire) d'une variable cible. Cette dernière est supposée sujette à un bruit blanc Gaussien, et, par conséquent, la vraisemblance des données a une correspondance directe avec les moindres carrés voir C. Bishop, 2007, par exemple. La sélection de variables a été, ces vingt dernières années, un des thèmes de recherche en statistique sur lequel le plus d'efforts ont été consacrés. Les solutions proposées s'appuient régulièrement sur la pénalisation en norme ℓ_1 du vecteur de régression, depuis les travaux précurseurs de Tibshirani, 1996 et l'introduction du Lasso (Least absolute skrinkage and selection operator, en anglais). En optimisation, le problème de départ considère généralement la log-vraisemblance du modèle linéaire, pénalisée par le nombre de variables actives. De nombreuses méthodes ont été développées dans ce cadre. Malheureusement, la pénalité associée, appelée pénalité ℓ_0, étant non connexe, le problème d'inférence devient donc combinatoire. La norme ℓ_1 peut alors être utilisée comme relaxation convexe. D'autres normes ad hoc et d'innombrables approches d'optimisation ont été introduites dans la littérature pour ce problème. Pour d'excellentes descriptions, nous conseillons Hastie, Tibshirani et Wainwright, 2015 ainsi que Bach, Jenatton et al., 2012. En parallèle, des développements Bayésiens ont été réalisés dans ce cadre, s'appuyant souvent sur les travaux de Mitchell et Beauchamp, 1988. Ainsi, une loi a priori de type spike and slab en anglais est employée. Certains coefficients se voient alors caractérisés par une loi piquée en 0, typiquement une distribution de Dirac. Les autres sont modélisés par une distribution, comme la distribution normale, plate, c'est-à-dire de variance a priori élevée. L'inférence consiste à allouer les variables aux deux types de lois.

Ce problème est combinatoire et rappelle la régression avec pénalité ℓ_0. Nous présentons ici une stratégie de modélisation et d'optimisation apportant un élément de solution.

5.2.1 Le modèle

Nous considérons un échantillon de n observations représentées par la matrice de régression X de taille $n \times p$ à coefficients réels ainsi que par le vecteur Y décrivant les n observations cibles réelles. Le modèle proposé dans Latouche, Mattei et al., 2016 s'écrit :

$$\begin{cases} Y &= X\beta + \varepsilon \\ \beta &= z \odot w. \end{cases}$$

Le symbole $\odot$ désigne le produit d'Hadamard, c'est-à-dire le produit terme à terme. Ainsi, le vecteur β à p éléments est tel que $\beta_j = z_j w_j$. De plus, le vecteur $z \in \{0,1\}^p$ est binaire alors que le vecteur w est supposé suivre une loi a priori normale $p(w|\alpha) = \mathcal{N}(w; 0, I_p/\alpha)$. Enfin, un bruit blanc Gaussien est retenu pour ε : $p(\varepsilon|\gamma) = \mathcal{N}(\varepsilon; 0, I_n/\gamma)$. La caractérisation de β permet alors de faire apparaître naturellement une loi a priori de type spike and slab (Mitchell et Beauchamp, 1988). En effet

$$p(\beta|z,\alpha) = \prod_{j=1}^{p} p(\beta_j|z_j,\alpha) = \prod_{j=1}^{p} \delta_0(\beta_j)^{1-z_j} \mathcal{N}(\beta_j; 0, 1/\alpha)^{z_j},$$

où $\delta_0(\cdot)$ désigne la distribution de Dirac centrée en 0. Un point clé de Latouche, Mattei et al., 2016 est que z n'est pas lui même caractérisé par une loi a priori de Bernoulli, contrairement à Mitchell et Beauchamp, 1988 et aux travaux récents en sélection de variables Bayésiennes. Le vecteur binaire est vu comme un paramètre à estimer, au même titre que α et γ. Comme nous le montrerons en section 5.2.3, le cadre Bayésien introduit ici est suffisant pour faire apparaître un critère auto-pénalisant sur le nombre de variables actives, sans qu'il soit nécessaire d'ajouter une couche supplémentaire dans le modèle hiérarchique.

L'objectif premier de ces travaux étant la sélection des variables dans le modèle de régression linéaire, nous proposons de considérer le problème de maximisation de la vraisemblance marginale $p(Y|X,z,\alpha,\gamma)$ par rapport à (z,α,γ), le vecteur β étant intégré. Cette optimisation est donc de type Bayésien empirique. Notons que la maximisation par rapport à z permet de se déplacer à travers l'espace des 2^p modèles Bayésiens, où un modèle est caractérisé par un ensemble particulier de variables actives. La sélection de variables est ainsi portée par le vecteur z. L'algorithme que nous avons développé s'appuie sur la loi a posteriori de w, sachant les données et (z,α,γ), qui a une forme explicite. Ci-dessous, nous notons $Q = ||z||_0$ le nombre de variables actives dans z et $Z = \mathrm{diag}(z)$ la matrice diagonale dont la diagonale est le vecteur z.

Proposition 5.1. *La loi a posteriori de w, sachant les données et (z,α,γ) est donnée par :*

$$p(w|Y,Z,\alpha,\gamma) = \mathcal{N}(w; m, S),$$

où $S = (\gamma Z X^T X Z + \alpha I_p)^{-1}$ et $m = \gamma S Z X^T Y$.

Notons que les supports de m et z coïncident presque surement et

$$m_z = \left(X_z^\intercal X_z + \frac{\alpha}{\gamma} I_p \right)^{-1} X_z^\intercal Y, \tag{5.1}$$

où une matrice (resp vecteur) indicée par z désigne la matrice (resp vecteur) restreinte à ses colonnes (resp éléments) pour lesquelles z est non nul. Ainsi, m_z correspond à l'estimateur ridge pour le modèle de régression dont les variables actives sont caractérisées par z.

5.2.2 Inférence

La vraisemblance marginale des données observées s'écrit :

$$p(Y|X,z,\alpha,\gamma) = \int_{\mathbb{R}^p} p(Y|X,w,z,\alpha,\gamma)p(w|\alpha)dw.$$

Le vecteur w ayant une loi a posteriori explicite comme indiqué en proposition 5.1, il est possible de dériver un algorithme EM afin de maximiser $p(Y|X,z,\alpha,\gamma)$. L'optimisation en z est singulière. En effet, chercher $\hat{z}$, un estimateur global de z, revient à tester l'ensemble des 2^p modèles Bayésiens possibles. Cette recherche exhaustive ne peut se faire que pour de faibles valeurs de p. Pour palier à ce problème, nous proposons une stratégie d'inférence à deux temps. Tout d'abord, une simple relaxation est employée afin de remplacer $z \in \{0,1\}^p$ par un vecteur $\tilde{z} \in [0,1]^p$. Un estimateur de $\tilde{z}$ est alors construit à partir de l'algorithme EM ci-dessous. Dans un second temps, la vraisemblance marginale est maximisée par rapport à $z \in \{0,1\}^p$ sur un chemin de modèles Bayésiens fabriqué à partir de l'estimateur de $\tilde{z}$ (section 5.2.3).

Pour simplifier les notations, Z désigne ci-dessous directement la matrice $\mathrm{diag}(\tilde{z})$ de diagonale $\tilde{z}$ utilisée par l'algorithme EM. L'étape E consiste à calculer l'espérance, par rapport à la loi a posteriori donnée par la proposition (5.1), de la log-vraisemblance des données complétées $\log p(Y,w|X,Z,\alpha,\gamma)$. Dans l'étape M, cette espérance est ensuite maximisée par rapport à $(\tilde{z},\alpha,\gamma)$ afin de produire des estimateurs.

Proposition 5.2. *En notant $\Sigma = S + mm^T$, l'espérance de la log-vraisemblance des données complétées est donnée par :*

$$\begin{aligned}\mathrm{E}_w[\log p(Y,w|Z,\alpha,\gamma)] = \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}Y^\intercal Y - \frac{\alpha}{2}\mathrm{Tr}(\Sigma) + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) \\ + \gamma\tilde{z}^\intercal(m \odot (X^\intercal Y)) - \frac{\gamma}{2}\tilde{z}^\intercal(X^\intercal X \odot \Sigma)\tilde{z}.\end{aligned}$$

Proposition 5.3. *Les estimateurs de α et γ maximisant l'espérance de la log-vraisemblance des données complétées (proposition 5.2) sont :*

$$\hat{\gamma}^{-1} = \frac{1}{n}\left\{Y^\intercal Y + \tilde{z}^\intercal(X^\intercal X \odot \Sigma)\tilde{z} - 2\tilde{z}^\intercal(m \odot (X^\intercal Y))\right\},$$

$$\hat{\alpha} = \frac{p}{\mathrm{Tr}(\Sigma)}.$$

En ce qui concerne l'estimation de $\tilde{z}$, un algorithme d'optimisation quadratique sous contraintes est utilisé :

$$\hat{\tilde{z}} = \underset{u\in[0,1]^p}{\arg\max}\left\{-\frac{1}{2}u^\intercal(X^\intercal X \odot \Sigma)u + u^\intercal(m \odot (X^\intercal Y))\right\}.$$

Les étapes E et M sont alternées jusqu'à convergence de la vraisemblance marginale des données.

5.2.3 Sélection de modèles

Aucune approximation asymptotique ou variationnelle n'est nécessaire, la log-vraisemblance marginale s'obtient de manière explicite.

Proposition 5.4. *À des constantes près, ne dépendant pas des paramètres, la log-vraisemblance marginale, également appelée log-vraisemblance de type II, peut s'écrire :*

$$\begin{aligned} -\log p(Y|z,\alpha,\gamma) &= -\log p(Y|m,z,\gamma) + \mathrm{pen}(z,\alpha,\gamma) \\ &= \frac{\gamma}{2}||Y - X_z m_z||_2^2 + \mathrm{pen}(z,\alpha,\gamma), \end{aligned}$$

où

$$\mathrm{pen}(z,\alpha,\gamma) = \frac{\alpha}{2}\left\|m\right\|_2^2 - \frac{\log\alpha}{2}\left\|m\right\|_0 - \frac{1}{2}\log\det(\gamma X_z^\intercal X_z + \alpha I_q) \text{ p.s.}$$

La pénalité $\mathrm{pen}(z,\alpha,\gamma)$ *correspond au facteur d'Occam en théorie Bayésienne.*

La proposition 5.4 fait donc apparaître l'erreur des moindres carrés, centrée sur l'estimateur ridge m_z associé au modèle caractérisé par z. Le terme de pénalité discuté ci-dessous pénalise la complexité du modèle. Optimiser z revient donc à chercher l'estimateur ridge minimisant cette fonction de perte des moindres carrés pénalisés.

LE RASOIR D'OCCAM En notant, $\lambda = (\alpha - \log\alpha)/2$ et $\kappa = \alpha/(\alpha - \log\alpha)$, la pénalité de la proposition 5.4 peut s'écrire :

$$\mathrm{pen}(z,\alpha,\gamma) = \lambda\left((1-\kappa)\left\|m\right\|_0 + \kappa\left\|m\right\|_2^2\right) - \frac{1}{2}\log\det(\gamma X_z^\intercal X_z + \alpha I_q).$$

Elle fait donc apparaître une combinaison convexe des pénalités ℓ_0 et ℓ_2. Ainsi, aussi bien le nombre de variables actives que l'amplitude des coefficients de régression sont contrôlés. Comme indiqué dans LATOUCHE, MATTEI et al., 2016, la pénalité de l'elastic net de ZOU et HASTIE, 2005, construite de manière ad hoc à partir des pénalités ℓ_1 et ℓ_2, peut donc être vue comme une relaxation convexe du rasoir d'Occam. Notons également que le rasoir d'Occam apparu dans la proposition 5.4 est fortement lié à la pénalisation du critère BIC. En effet, en considérant une loi a priori plate pour w et si S est de rang plein, alors la pénalité peut être approchée par $-(1/2)q\log n$.

CHEMIN DE MODÈLES L'algorithme EM de la section 5.2.2 permet de construire des estimateurs de $\tilde{z} \in [0,1]^p$, α, et γ. En pratique, nous cherchons ici à sélectionner un modèle Bayésien, c'est-à-dire un ensemble de variables. Un vecteur $z \in \{0,1\}^p$ doit donc être déterminé. Partant d'aucune variable active, l'heuristique proposée dans LATOUCHE, MATTEI et al., 2016 consiste à ajouter des variables au fur et à mesure dans un ordre fixé par l'estimateur de $\tilde{z}$. Ainsi, la variable ayant le coefficient $\tilde{z}_j$ le plus élevé est ajoutée en premier, suivie de la variable ayant le 2ème coefficient le plus élevé, etc ... La vraisemblance marginale de la proposition 5.4 est alors calculée sur ce chemin de modèles et le modèle maximisant la marginale est retenu comme estimateur de z.

À titre illustratif, la figure 5.1 montre le comportement sur des données simulées de la vraisemblance marginale lorsque les variables sont ajoutées. Le graphique de gauche présente les termes $\tilde{z}_j$ ordonnés alors que celui de droite donne les valeurs de la vraisemblance marginale pour les différents modèles considérés. Cette dernière a un pique pour 5 variables ce qui correspond exactement au nombre de variables utilisées pour la simulation.

5.2.4 Expérimentations numériques

Nous présentons maintenant une partie des simulations de LATOUCHE, MATTEI et al., 2016 utilisées pour tester la méthodologie d'inférence. Ainsi, les vecteurs w et ε sont tirés à partir

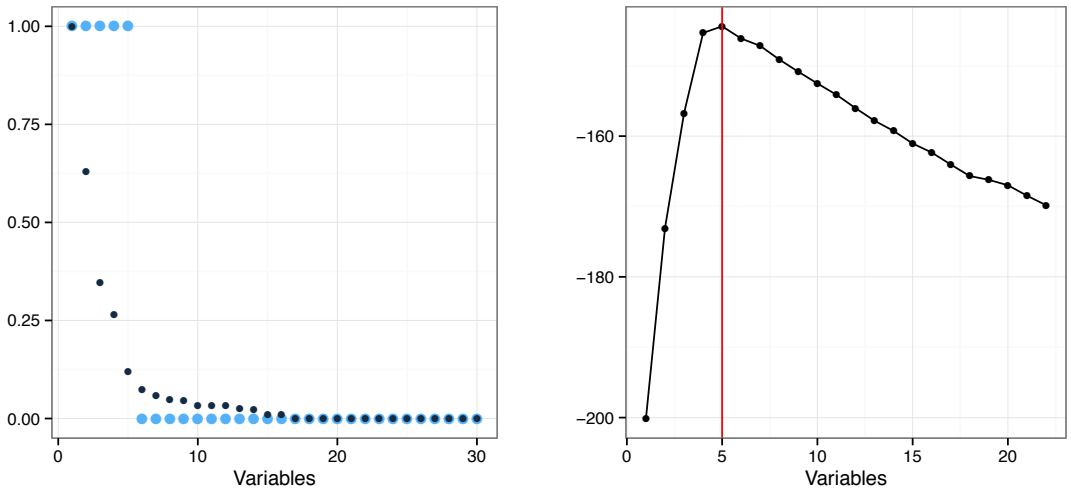


Figure 5.1 – Graphique de gauche : coefficients $\tilde{z}_j$ ordonnés (bleu foncé). Vraies valeurs (binaires) de z (bleu clair). Valeur de la log-vraisemblance marginale à droite. En rouge, le véritable nombre de variables utilisé pour cette simulation.

des lois $\mathcal{N}(0, I_p/\alpha)$ et $\mathcal{N}(0, I_n/\gamma)$, respectivement. La matrice X de régression est elle simulée selon une loi normale centrée et de covariance R. Seuls les résultats pour des matrices R de Toeplitz sont ici donnés. Ainsi la structure de corrélation est la suivante :

$$\begin{cases} R_{ii} & = 1, \forall i = 1, ...p \\ R_{ij} & = \rho^{|i-j|}, \forall i \neq j = 1, \ldots, p. \end{cases}$$

Le vecteur z est alors construit en choisissant uniformément au hasard q variables actives parmi p. Enfin, 100 jeux de données sont simulés pour trois tailles ($n = p/2, n = p, n = 2p$), et $\rho = 0.25$, $p = 100$, $q = 40$, $\alpha = 1$, ainsi que $\gamma = 1$. Notre approche, appelée spiny ci-dessous, est comparée avec d'autres méthodes de régression parcimonieuse : le lasso de Tibshirani, 1996, le lasso adaptatif (adalasso) de Zou, 2006, la procédure de stabilité de sélection (stabsel) de Meinshausen et Bühlmann, 2010, l'algorithme clere (Yengo, Jacques et al., 2016), et la procédure de propagation d'espérance (ssep) de D. Hernández-Lobato, J. Hernández-Lobato et Dupont, 2013. Les résultats sont évalués à l'aide de trois critères : l'erreur moyenne des moindres carrés, le F-score (moyenne harmonique entre la précision et le rappel), et le nombre estimé de variables actives ($\hat{q}$). Ils sont présentés en figure 5.2. Ainsi, la procédure de sélection de modèles que nous proposons, associée à un algorithme EM, a des performances compétitives par rapport aux autres méthodes étudiées. Elle permet en particulier de construire un estimateur robuste du nombre de variables actives, contrairement au lasso par exemple qui a tendance à surestimer le nombre de variables.

5.3 ACP globalement parcimonieuse

L'analyse en composantes principales (ACP) (Pearson, 1901 ; Hotelling, 1933) est une des méthodes les plus utilisées pour pré-traiter les données et réduire leur dimension. Le principe consiste à projeter les données sur un sous-espace généré à partir des vecteurs propres principaux de la matrice de covariance empirique. Il a été montré dans Roweis, 1998 ; Tipping et C.M. Bishop, 1999 que ce sous espace pouvait être obtenu à partir d'un estimateur

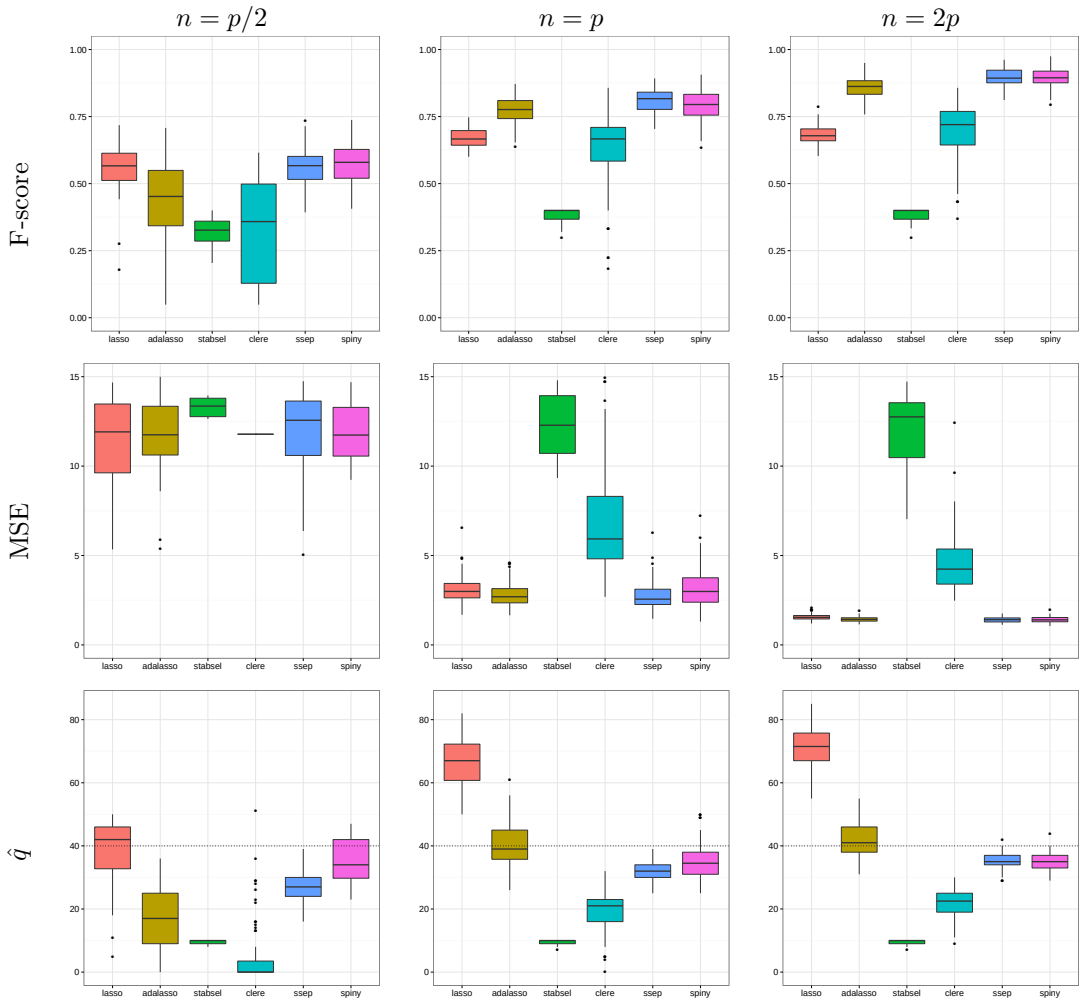


Figure 5.2 – Boîtes à moustaches présentant les résultats de six méthodes de régression parcimonieuse sur 100 simulations à chaque fois, en terme de F-score, d'erreur moyenne des moindres carrés (MSE), et du nombre estimé de variables actives ($\hat{q}$).

du maximum de vraisemblance d'un modèle probabiliste factoriel particulier, appelé ACP probabiliste (ACPP). Une des limites de l'ACP est que les nouvelles variables, obtenues après projection et appelées composantes principales, sont des combinaisons linéaires de l'intégralité des variables de départ. Dans le cas de données de grande dimension, cela complique considérablement l'interprétabilité des résultats. Dès lors, de nombreuses techniques ont été proposées dans la littérature afin de projeter les données sur un sous-espace généré à partir de vecteurs parcimonieux, c'est-à-dire ayant des composantes fixées exactement à zéro. Elles considèrent au départ pour la plupart un problème d'optimisation où le nombre de paramètres actifs, dans les vecteurs de projection, est pénalisé. Elles introduisent ensuite des relaxations convexes ou partiellement convexes afin d'éviter la combinatoire associée. À titre d'exemple, nous pouvons citer l'approche de Zou, Hastie et Tibshirani, 2006, basée sur une pénalisation de type ℓ_1. D'autres méthodes, comme celle d'Archambeau et

Bach, 2009, considèrent un cadre Bayésien et s'appuient sur des lois a priori parcimonieuses. Malheureusement, toutes ces approches ne forcent pas les vecteurs de projection à avoir le même support. Par conséquent, les composantes principales obtenues doivent toutes être interprétées séparément. Cela ne pose pas de problème critique si l'objectif est de visualiser les données dans un espace de plus faible dimension. En revanche, si le praticien cherche à savoir quelles sont, dans son jeu de données, les variables globalement les plus importantes, alors la notion de parcimonie doit être redéfinie. Dans Bouveyron, Latouche et Mattei, 2017a, nous avons proposé un cadre méthodologique permettant de faire de l'ACP globalement parcimonieuse, c'est-à-dire de chercher des vecteurs de projection parcimonieux, partageant le même support.

5.3.1 Modèle avec bruit

Un échantillon de n observations iid $x_1, \ldots, x_n \in \mathbb{R}^p$ est considéré. Pour simplifier les notations, ces dernières sont stockées dans une matrice X de taille $n \times p$ où la ligne i de X correspond à $x_i^\intercal$. Le modèle ACPP (Roweis, 1998 ; Tipping et C.M. Bishop, 1999) s'écrit :

$$x = Wy + \varepsilon, \tag{5.2}$$

où $y \sim \mathcal{N}(0, I_d)$ est un vecteur de dimension d ($<= p$) supposé Gaussien, W est une matrice de taille $p \times d$, appelée loading matrix en anglais, et $\varepsilon|\sigma \sim \mathcal{N}(0, \sigma^2 I_p)$ est un terme de bruit blanc Gaussien. Dans ce cadre, il apparaît que les composantes principales de l'ACP classique sont fortement liées à l'estimateur W_{MV} du maximum de vraisemblance de la matrice W. En effet, soient A la matrice de taille $p \times d$ des vecteurs propres, ordonnés en colonne, de la matrice de covariance empirique des données, et Λ la matrice diagonale de taille $d \times d$ des valeurs propres associées, alors

$$W_{\text{MV}} = A(\Lambda - \sigma^2 I_d)^{1/2} R,$$

où R est une matrice orthogonale arbitraire.

Afin de sélectionner les variables de départ dans les observations x_i, des lignes complètes de W doivent être mises à zéro. Ainsi, en notant $v \in \{0, 1\}^p$ le vecteur binaire caractérisant les variables actives, à l'image des travaux présentés à la section précédente, et $V = \text{diag}(v)$, nous remplaçons le modèle ACPP par :

$$x = VWy + \varepsilon,$$

où W, y, et ε sont inchangés. Les 0 dans le vecteur v impliquent des lignes nulles dans la matrice VW. L'inférence de VW impliquera donc que le sous-espace de projection des données sera généré par des vecteurs partageant le même support, c'est-à-dire les mêmes éléments non-nuls.

Cadre Bayésien Le modèle avec bruit étant maintenant introduit, l'objectif est de réaliser son inférence en pénalisant le nombre de variables actives. Nous employons ici une stratégie d'inférence Bayésienne inspirée de nos travaux en régression parcimonieuse, présentés à la section précédente. Ainsi, les coefficients W_{ij} sont supposés iid a priori et modélisés par la loi normale : $W_{ij}|\alpha \sim \mathcal{N}(0, 1/\alpha^2)$. La loi marginale des données observées est alors donnée par le théorème 5.1.

Théorème 5.1. *La loi marginale des données observées (sous l'hypothèse d'indépendance conditionnelle) dans le modèle avec bruit s'écrit :*

$$p(X|v,\alpha,\sigma) = \prod_{i=1}^{n} p(x_i|v,\alpha,\sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^{p\times d}} p(x_i|W,v,\alpha,\sigma)p(W)dW,$$

où

$$p(x|v,\alpha,\sigma) = \frac{e^{-\frac{||x_{\bar{v}}||_2^2}{2\sigma^2}}}{(2\pi)^{p/2}\sigma^{p-q}}||x_v||_2^{1-q/2} \int_0^\infty \frac{u^{q/2}e^{-\sigma^2u^2}}{(1+(u/\alpha)^2)^{d/2}} J_{q/2-1}(u||x_v||_2)du.$$

Le vecteur $\bar{v} = 1 - v$ désigne le complémentaire de v et $x_{\bar{v}}$ est le vecteur x restreint à ses composantes actives (1) dans $\bar{v}$. $J_{q/2-1}(\cdot)$ est la fonction de Bessel de première espèce, d'ordre $q/2-1$.

La densité de x sachant v, α, et σ fait intervenir une intégrale unidimensionnelle qui malheureusement est de type Hankel et par conséquent difficile à déterminer, même numériquement. L'optimisation de la marginale du modèle avec bruit soulève donc de nombreux problèmes techniques. Dans Bouveyron, Latouche et Mattei, 2017a, nous avons proposé de remplacer ce modèle en nous appuyant sur la version sans bruit du modèle ACPP.

5.3.2 Modèle sans bruit

Afin d'obtenir une expression exacte et calculable pour la vraisemblance marginale des données, nous avons modifié le modèle avec bruit. Ainsi, un modèle ACPP sans bruit est retenu pour les variables actives alors que les variables inactives sont supposées générées par un bruit blanc Gaussien $\varepsilon_1|\sigma_1 \sim \mathcal{N}(0,\sigma_1^2 I_p)$. Plus précisément,

$$x = VWy + \bar{V}\varepsilon_1 + V\varepsilon_2,$$

où $\bar{V} = \text{diag}(\bar{v})$ et $\varepsilon_2|\sigma_2 \sim \mathcal{N}(0,\sigma_2^2 I_p)$. Nous nous intéressons au cas $\sigma_2 \to 0$. La définition d'une loi symétrique multivariée de Bessel est rappelée ci-dessous. Nous donnons ensuite l'expression de la vraisemblance marginale.

Definition 5.1. *Un vecteur aléatoire suit une distribution symétrique multivariée de Bessel de paramètres $\beta > 0$ et $\nu > -k/2$ si sa distribution s'écrit :*

$$\forall z \in \mathbb{R}^k,\ \text{Bessel}(z|\beta,\nu) = \frac{2^{-k-\nu+1}\beta^{-k-\nu}}{\Gamma(\nu+k/2)\pi^{k/2}}||z||_2^\nu K_\nu(||z||_2/\beta).$$

$K_\nu(\cdot)$ est la fonction de Bessel modifiée (voir par exemple Abramowitz et Stegun, 1965, chapitres 10 et 11) de second espèce, d'ordre ν.

Théorème 5.2. *Dans le cas limite sans bruit pour les variables actives ($\sigma_2 \to 0$), x converge en probabilité vers un vecteur aléatoire dont la distribution est :*

$$p(x|v,\alpha,\sigma_1^2) = \mathcal{N}(x_{\bar{v}}|0,\sigma_1 I_{p-q})\text{Bessel}(x_v|1/\alpha,(d-q)/2). \tag{5.3}$$

La log-vraisemblance marginale associée (sous l'hypothèse d'indépendance conditionnelle) est donc $\log p(X|v,\alpha,\sigma_1^2) = \sum_{i=1}^n \log p(x_i|v,\alpha,\sigma_1^2)$ où $p(x_i|v,\alpha,\sigma_1^2)$ est obtenue par (5.3). Plusieurs choix, discutés dans Bouveyron, Latouche et Mattei, 2017a, sont possibles pour choisir l'hyperparamètre σ_1, à v fixé. Par exemple, un estimateur peut être construit en calculant l'erreur standard des variables non sélectionnées dans v. De plus, à v fixé, la fonction $\alpha \mapsto \log p(X|v,\alpha,\sigma_1^2)$ est convexe en α. Dans ce cadre, une valeur optimale peut donc être trouvée pour α en s'appuyant sur un algorithme simple d'optimisation convexe.

5.3.3 Inférence

L'optimisation de la vraisemblance marginale $p(X|v, \alpha, \sigma_1^2)$ par rapport à v est un problème combinatoire. Afin de proposer un cadre d'estimation, nous nous appuyons sur les travaux présentés en section 5.2. Ainsi, un schéma à deux temps est considéré. Une simple relaxation est utilisée pour remplacer $v \in \{0,1\}^p$ par $u \in [0,1]^p$. Une fois un estimateur de u obtenu, la vraisemblance marginale est alors optimisée sur un chemin de modèles. Malheureusement, le cadre général de l'ACP probabiliste pose ici deux contraintes majeures. Tout d'abord, bien que permettant d'introduire une expression calculable pour la vraisemblance marginale, le modèle sans bruit n'est pas adapté pour l'inférence de u. C'est en effet un modèle théorique construit à partir d'un cas limite $\sigma_2 \to 0$ et l'expression (5.3) ne tient que pour un vecteur v binaire. Pour l'inférence, le modèle avec bruit, et relaxation, est donc employé :

$$x = UWy + \varepsilon,$$

où $U = \mathrm{diag}(u)$. Ensuite, même dans le cas du modèle avec bruit, un algorithme EM ne peut être dérivé. Effectivement, en notant $\theta = (u, \alpha, \sigma)$ l'ensemble des paramètres du modèle, la loi a posteriori $p(W, Y|X, \theta)$ n'a pas de forme analytique. Par conséquent, un algorithme VEM est utilisé. Ainsi, en notant Y la matrice de taille $n \times d$ où la ligne i correspond au vecteur $y_i^\intercal$, une borne variationnelle est optimisée par rapport à la loi $r(Y, W) = r(Y)r(W)$ factorisée et les paramètres dans θ. Les étapes de maximisation sont alternées jusqu'à convergence de la borne. Les deux propositions suivantes donnent les étapes E et M de l'algorithme variationnel.

Proposition 5.5. *Les lois $r(Y)$ et $r(W)$, maximisant la borne inférieure de l'algorithme VEM pour le modèle avec bruit et relaxation, sont données par :*

$$r(Y) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mu_i, \Sigma), \tag{5.4}$$

et

$$r(W) = \prod_{k=1}^{p} \mathcal{N}(w_k|m_k, S_k), \tag{5.5}$$

où, $\forall i \in \{1, ..., n\}$ et $\forall k \in \{1, ..., p\}$:

$$\mu_i = \frac{1}{\sigma^2}\Sigma M^\intercal U x_i,\ m_k = \frac{u_k}{\sigma^2} S_k \sum_{i=1}^{n} x_{ik}\mu_i,$$

$$\Sigma^{-1} = I_d + \frac{1}{\sigma^2} M^\intercal U^2 M + \frac{1}{\sigma^2}\sum_{k=1}^{p} u_k^2 S_k,\ S_k^{-1} = \alpha^2 I_d + \frac{n u_k^2}{\sigma^2}\Sigma + \frac{u_k^2}{\sigma^2}\mathcal{M}^\intercal\mathcal{M},$$

$$M = (m_1, ...m_p)^\intercal \ \ et \ \ \mathcal{M} = (\mu_1, ...\mu_n)^\intercal.$$

Proposition 5.6. *Les estimateurs des paramètres dans θ, maximisant la borne inférieure de l'algorithme VEM pour le modèle avec bruit et relaxation, sont donnés par :*

$$\alpha^* = \left(\frac{1}{dp}\sum_{k=1}^{p} \mathrm{Tr}(S_k + m_k m_k^T)\right)^{-1/2}, \tag{5.6}$$

$$\sigma^* = \sqrt{\frac{\mathrm{Tr}(XX^T + XUM\mathcal{M})}{np} + \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{p} u_k^2 \mathrm{Tr}[(\Sigma + \mu_i\mu_i^T)(S_k + m_i m_i^T)]}, \tag{5.7}$$

et, pour $k \in \{1, ..., p\}$,

$$u_k^* = \mathrm{argmin}_{u\in[0,1]} \frac{u^2}{2\sigma^2}\sum_{i=1}^{n} \mathrm{Tr}[(\Sigma + \mu_i\mu_i^T)(S_k + m_i m_i^T)] - u\sum_{i=1}^{n} x_{ik} m_k^T \mu_i. \tag{5.8}$$

CHEMIN DE MODÈLES À l'image de LATOUCHE, MATTEI et al., 2016, l'algorithme VEM permet de construire un estimateur $\hat{u}$ de $u \in [0,1]^p$. L'objectif fondamental est ici d'estimer le nombre de variables actives. Un vecteur $v \in \{0,1\}^p$ doit donc être retenu. Partant d'un modèle sans aucune variable active, les variables sont ajoutées au fur et à mesure dans un ordre fixé par l'estimateur $\hat{u}$. Ainsi, la variable ayant le plus grand terme $\hat{u}_j$ est d'abord employée, suivie de la variable ayant le 2ème terme le plus élevé, etc ... La vraisemblance marginale du modèle sans bruit, donnée par (5.3), est alors calculée sur ce chemin de modèles. Le vecteur binaire $\hat{v}$ maximisant cette dernière est alors retenu comme estimateur de v.

La figure 5.3 illustre cette heuristique d'optimisation sur un jeu de données simulées. Le graphique de gauche indique les $\hat{u}_j$ ordonnés, dans l'ordre décroissant, alors que celui de droite donne les valeurs associées de la vraisemblance marginale du modèle sans bruit. Cette dernière atteint son maximum pour 10 variables qui est bien le nombre de variables utilisées dans la simulation.

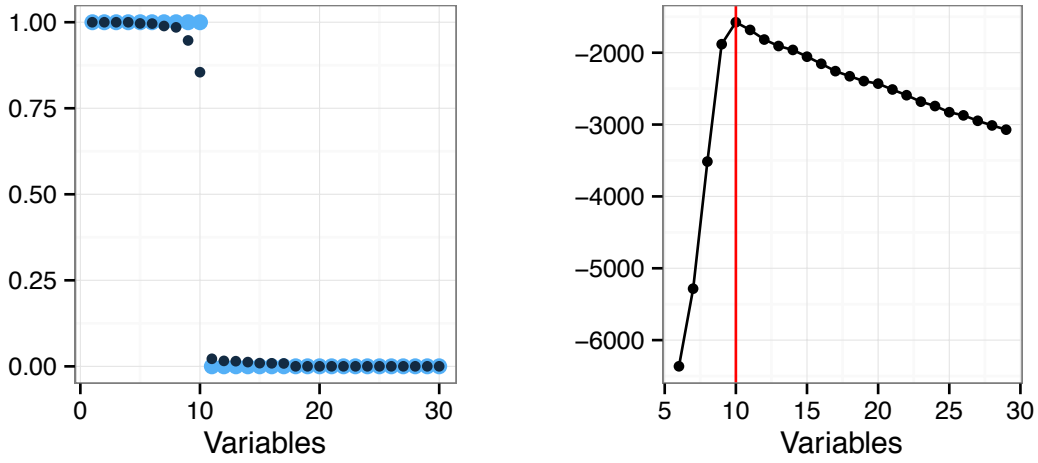


Figure 5.3 – Graphique de gauche : coefficients $\hat{u}_j$ ordonnés (bleu foncé). Vraies valeurs (binaires) de v (bleu clair). Valeur de la vraisemblance marginale du modèle sans bruit à droite. En rouge, le véritable nombre de variables utilisées pour cette simulation.

5.3.4 Expérimentations numériques

Une partie seulement des expériences sur simulations numériques de BOUVEYRON, LATOUCHE et MATTEI, 2017a sont ici présentées. Tout d'abord, n observations $(z_1, \ldots, z_n)$ sont tirées de manière *iid* à partir d'une loi normale $\mathcal{N}(0, R)$ où $R = \mathrm{diag}(R^1, \ldots, R^4)$ est une matrice symétrique diagonale par bloc, telle que le bloc R^ℓ est donné par $R^\ell_{ii} = 0.3$ et

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
ACPS	20.7 ± 0.7	21.2 ± 0.7	21.5 ± 0.7	21.7 ± 0.5	25.2 ± 2.1
ACPSS	66.7 ± 21.4	71.5 ± 20	86.7 ± 14.2	95.6 ± 8.9	98.2 ± 7.2
ACPPGP	$\mathbf{86.8 \pm 7.06}$	$\mathbf{93.9 \pm 3.66}$	$\mathbf{97.2 \pm 2.55}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{1 \pm 0}$

Table 5.1 F-score moyen (et écart type) $\times 100$ pour la sélection de modèles avec bruit Gaussien.

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
ACPS	20.8 ± 0.6	21.3 ± 0.6	21.6 ± 0.8	21.8 ± 0.6	25.3 ± 1.7
ACPSS	60.6 ± 22.4	63.9 ± 25.2	82.7 ± 18.1	$\mathbf{94.2 \pm 10.2}$	97.4 ± 9.5
ACPPGP	$\mathbf{74.2 \pm 10}$	$\mathbf{77.6 \pm 9.09}$	$\mathbf{79.7 \pm 8.38}$	88 ± 5.95	$\mathbf{99.2 \pm 1.4}$

Table 5.2 F-score moyen (et écart type) $\times 100$ pour la sélection de modèles avec bruit de Laplace.

$R^{\ell}_{ij} = \rho, \forall i, j = 1, \ldots, p/4, i \neq j$. Ensuite, l'estimateur non parcimonieux W_{MV} du modèle d'ACPP est construit à partir de ces données. Finalement, pour un vecteur binaire v donné et une matrice $V = \mathrm{diag}(v)$, les observations à analyser sont générées à partir d'un modèle avec bruit :

$$x_i = VW_{\mathrm{MV}} y_i + \varepsilon_i, \forall i.$$

Comme dans le modèle d'ACPP, les vecteurs $y_1, \ldots, y_n$ sont simulés selon une loi normale centrée et réduite. En ce qui concerne le terme de bruit, deux scénarios sont considérés. Le premier avec bruit Gaussien et le second avec bruit de Laplace, tous les deux étant centrés et réduits. Nous avons fixé $p = 200$, $d = 10$, et $q = 20$ le nombre de variables actives. Cinq situations sont alors testées, suivant la taille des données : $n = p/5$, $p/4$, $n = \lfloor p/3 \rfloor$, $n = p/2$ ainsi que $n = p$. Toutes les expériences sont répétées 50 fois et les performances sont évaluées à partir du F-score (voir section précédente). Notre approche, appelée ACPP globalement parcimonieuse (ACPPGP) ci-dessous, est comparée avec les algorithmes de Zou, Hastie et Tibshirani, 2006 (ACPS) et Jenatton, Obozinski et Bach, 2009 (ACPSS). Seuls les résultats en sélection de modèles, c'est-à-dire sur l'estimation de v, sont présentés en figures 5.1 et 5.2. Sur ces exemples, ACPPGP obtient toujours le meilleur F-score, sauf dans un cas.

5.4 Sélection de la dimension en ACP Bayésienne

Introduite par Pearson, 1901 et redécouverte par Hotelling, 1933, l'ACP est aujourd'hui utilisée de manière quasi systématique dans de très nombreux champs d'applications. Cependant, force est de constater qu'aucune méthode de sélection du nombre de vecteurs propres principaux, utilisés pour la projection des données, n'est reconnue comme méthode de référence. Une heuristique simple consiste à regarder la décroissance des valeurs propres de la matrice de covariance empirique. Cette technique ad hoc, popularisée par les travaux de Cattell, 1966, a été modifiée et améliorée de nombreuses fois dans la littérature. En parallèle, des cadres méthodologiques différents ont été employés pour répondre à cette question, des tests d'hypothèses (Jolliffe, 2002) jusqu'aux approches probabilistes basées sur le modèle d'ACP probabiliste (ACPP) de Tipping et C.M. Bishop, 1999, en passant par la validation croisée (Wold, 1978 ; Josse et Husson, 2012). Ces stratégies types d'estimation sont pour la plupart basées sur des considérations asymptotiques sur le nombre d'observations. Or, dans ce cadre, il a été montré récemment qu'un seuillage simple des valeurs propres était suffisant pour estimer la dimension de l'espace de projection (Gavish et Do-

NOHO, 2014). Des critères de sélection de modèles, souvent construits à partir de stratégies d'inférence Bayésienne, ont donc été introduits. Malheureusement, les outils associés, utilisés pour l'estimation sur données réelles, ont des coûts algorithmiques prohibitifs (voir HOFF, 2007, par exemple). Dans BOUVEYRON, LATOUCHE et MATTEI, 2017b, nous avons donc proposé un cadre méthodologique permettant d'apporter une solution à ce problème.

5.4.1 Cadre

À nouveau, le modèle d'ACPP de TIPPING et C.M. BISHOP, 1999 est retenu comme point de départ pour la dérivation d'une stratégie d'inférence. Ainsi, pour toutes les n observations iid x_i de l'échantillon, nous avons :

$$x_i = W y_i + \varepsilon_i,$$

où $y_i \sim \mathcal{N}(0, I_d)$ et $\varepsilon_i|\sigma \sim \mathcal{N}(0, \sigma^2 I_p)$. Pour plus de détails, en particulier sur les liens existants entre l'ACP classique et le modèle ACPP, le lecteur est renvoyé au début de la section 5.3.1. L'objectif est ici d'estimer d à partir des données. Le cadre considéré pour l'instant est suffisant pour construire un critère de sélection de modèles de type vraisemblance pénalisée. En effet, en s'appuyant sur les estimateurs du maximum de vraisemblance du modèle ACPP, d peut être choisi de la manière suivante :

$$d^* \in \text{argmax}_{d \in \{1,\dots,p-1\}} \{\log p(X|W_{\text{ML}}, \sigma_{\text{ML}}, \mathcal{M}_d) - \text{pen}(d)\}.$$

Comme précédemment, la matrice X contient les vecteurs x_i de $\mathbb{R}^p$ et est de taille $n \times p$. De plus, $\mathcal{M}_d$ désigne le modèle ACPP à d fixé, et $(W_{\text{ML}}, \sigma_{\text{ML}})$ les estimateurs associés. Enfin, pen(d) est une pénalité qui augmente avec d. De nombreux choix sont possibles pour cette dernière pouvant donner naissance par exemple à un critère AIC (AKAIKE, 1974) ou BIC (SCHWARZ, 1978). Comme discuté dans l'introduction de cette section, les résultats théoriques concernant la qualité de l'estimation de d à partir de ce type de stratégies ne sont valides que dans un contexte asymptotique. Pour palier à cette limite, nous avons montré dans BOUVEYRON, LATOUCHE et MATTEI, 2017b qu'une structure a priori particulière pour ACPP permettait d'obtenir une forme analytique pour la vraisemblance marginale. Cette vraisemblance peut alors être utilisée dans ce cadre Bayésien pour estimer la valeur de d.

Ainsi, comme dans BOUVEYRON, LATOUCHE et MATTEI, 2017a, les coefficients W_{ij} de la matrice W sont supposés indépendants a priori et caractérisés par une loi normale : $W_{ij}|\phi \sim \mathcal{N}(0, \phi^{-1})$. Notons que le terme de variance est en ϕ^{-1} cette fois ci, et non pas en ϕ^{-2}, pour des raisons techniques. Une vraisemblance marginale peut alors être calculée (voir théorème 5.2) dans un cas limite où la variance des termes de bruit tend vers 0. Or cette variance représente l'information non expliquée par les nouvelles données y_i dans $\mathbb{R}^d$. L'objectif étant d'estimer d justement, cette dernière doit être retenue pour l'estimation. La vraisemblance marginale de BOUVEYRON, LATOUCHE et MATTEI, 2017a n'est donc pas adaptée à ce problème. Une loi gamma a priori est alors considérée : $\sigma^2 \sim \text{Gam}(a, b)$. Pour un choix particulier des hyperparamètres (b, ϕ), le théorème suivant montre que la vraisemblance marginale associée s'écrit exactement et se calcule facilement.

Théorème 5.3. *Soit $d \in \{1, \dots, d\}$. Sous la loi a priori normale gamma considérée avec $b = \phi/2$, la log-vraisemblance marginale (sous l'hypothèse d'indépendance conditionnelle) du*

modèle $\mathcal{M}_d$ d'ACPP à d fixé est donnée par :

$$\begin{aligned}\log p(X|a,\phi,\mathcal{M}_d) &= \sum_{i=1}^{n} \log p(x_i|a,\phi,\mathcal{M}_d) \\ &= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(2\phi^{-1}) - n\log\Gamma(a+d/2) \\ &\quad + (a+\frac{d-p}{2})\sum_{i=1}^{n}\log(\frac{\sqrt{\phi}||x_i||_2}{2}) + \sum_{i=1}^{n}\log K_{a+(d-p)/2}(\sqrt{\phi}||x_i||_2),\end{aligned}$$

où $K_{a+(d-p)/2}(\cdot)$ est la fonction de Bessel modifiée de second espèce, d'ordre $a+(d-p)/2$.

La preuve de ce théorème est donnée dans Bouveyron, Latouche et Mattei, 2017b. Elle s'appuie sur des lois de Laplace généralisées dont nous rappelons la définition ci-dessous ainsi que sur les propositions suivantes.

Definition 5.2. *Un vecteur aléatoire $z \in \mathbb{R}^p$ suit une loi multivariée asymétrique généralisée de Laplace $\mathrm{GAL}_p(\Sigma,\mu,s)$ de paramètres $s>0, \mu\in\mathbb{R}^p$ et $\Sigma \in \mathcal{S}_p^+$ si sa fonction caractéristique est :*

$$\forall u \in \mathbb{R}^p,\ \phi_{\mathrm{GAL}_p(\Sigma,\mu,s)}(u) = \left(\frac{1}{1+\frac{1}{2}u^\intercal\Sigma u - i\mu^\intercal u}\right)^s.$$

Proposition 5.7. *Si $z|\Sigma,\mu,s \sim \mathrm{GAL}_p(\Sigma,\mu,s)$, nous avons $\mathrm{E}[z]=s\mu$ et $\mathrm{V}(z)=s(\Sigma+\mu\mu^\intercal)$. De plus, si Σ est définie positive, la densité de z est donnée par :*

$$\forall x\in\mathbb{R}^p,\ f_z(x) = \frac{2e^{\mu^\intercal\Sigma^{-1}x}}{(2\pi)^{p/2}\Gamma(s)\sqrt{\det\Sigma}}\left(\frac{Q_\Sigma(x)}{C(\Sigma,\mu)}\right)^{s-p/2} K_{s-p/2}\left(Q_\Sigma(x)C(\Sigma,\mu)\right),$$

où $Q_\Sigma(x)=\sqrt{x^\intercal\Sigma^{-1}x}$ et $C(\Sigma,\mu)=\sqrt{2+\mu^\intercal\Sigma^{-1}\mu}$.

Proposition 5.8. *Soient $s_1,s_2>0, \mu\in\mathbb{R}^p$ et $\Sigma\in\mathcal{S}_p^+$. Si $z_1|\Sigma,\mu,s_1 \sim \mathrm{GAL}_p(\Sigma,\mu,s_1)$ et $z_2|\Sigma,\mu,s_2\sim\mathrm{GAL}_p(\Sigma,\mu,s_2)$ sont des vecteurs aléatoires indépendants, alors :*

$$z_1+z_2|\Sigma,\mu,s_1,s_2 \sim \mathrm{GAL}_p(\Sigma,\mu,s_1+s_2).$$

Proposition 5.9 (Gauss-Laplace)**.** *Soient $s>0$ et $\Sigma\in\mathcal{S}_p^+$. Si $u|s\sim\mathrm{Gamma}(s,1)$ et $x|\Sigma\sim\mathcal{N}(0,\Sigma)$ sont indépendants de u, alors :*

$$\sqrt{u}x|\Sigma,s\sim\mathrm{GAL}_p(\Sigma,0,s).$$

5.4.2 Le choix des hyperparamètres

Nous avons montré précédemment que fixer $b=\phi/2$ était suffisant pour obtenir une forme analytique de la vraisemblance marginale. Deux hyperparamètres doivent donc être choisis : le paramètre de forme a de la loi a priori gamma ainsi que le paramètre de précision ϕ de la loi normale. Une première observation est qu'à mesure que d augmente, il est attendu que la variance du bruit diminue. Cette caractéristique peut être intégrée en centrant la loi gamma sur un estimateur $\hat{\sigma}$ de σ. Ainsi, a est choisi de sorte que $\mathrm{E}(\sigma)\propto\hat{\sigma}$, pour chaque d testé. Plus précisément, afin que ϕ puisse avoir un rôle double sur le contrôle de W ainsi que sur la variance du bruit, nous avons fixé $a=\hat{\sigma}/\phi$. Plusieurs choix, discutés dans Bouveyron, Latouche et Mattei, 2017b, sont possibles pour l'estimation de σ. Dans

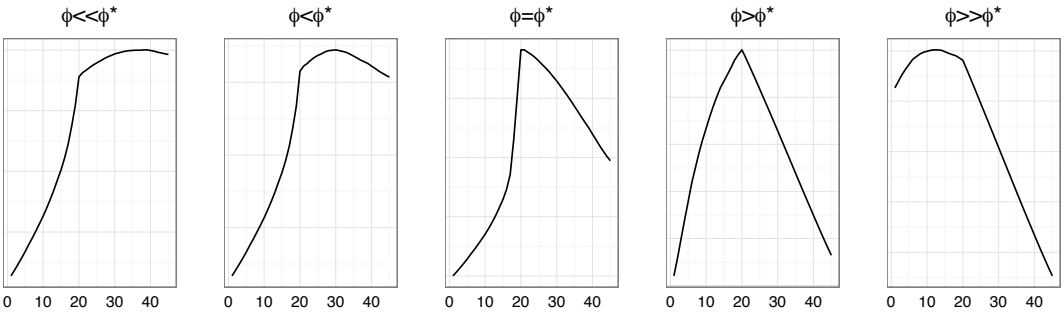


Figure 5.4 – La log-vraisemblance marginale en fonction de d, pour des valeurs de ϕ différentes. ϕ^* correspond à la valeur retenue par l'heuristique. Résultats obtenus sur données simulées avec $d = 20$.

toutes les expériences que nous avons réalisées, nous nous sommes appuyés sur l'estimateur du maximum de vraisemblance du modèle ACPP.

En ce qui concerne la détermination de ϕ, une heuristique est utilisée, basée sur un constat élémentaire. La vraisemblance marginale en fonction de d doit avoir deux phases successives. Une première phase tout d'abord où l'augmentation de d induit une augmentation sensible de la vraisemblance, les données construites expliquant de plus en plus d'information (signal). Une seconde phase alors où les dimensions ajoutées correspondent à du bruit et où la vraisemblance diminue. Cette décroissance est clé. En effet, dans des applications réelles, il est préférable de surestimer d que l'inverse. Par conséquent, ϕ est choisi de manière à ce que la pente (gain) avant le maximum soit supérieure à la pente (perte) après le maximum. À titre illustratif, la figure 5.4 donne le comportement de la log-vraisemblance marginale en fonction de d sur des données simulées, pour des choix de ϕ différents.

5.4.3 Expérimentations numériques

Nous présentons ici une partie des expériences de Bouveyron, Latouche et Mattei, 2017b sur données simulées. Afin de nous comparer à des valeurs de référence, nous considérons le modèle contraint d'ACPP introduit dans Bouveyron, Celeux et Girard, 2011. Ce dernier fait l'hypothèse que la matrice de variance covariance des données dans X a seulement deux valeurs propres r et s. Le rapport signal sur bruit s'écrit alors simplement :

$$\text{SNR} = \frac{rd}{p-d}.$$

Dans nos simulations, nous avons fixé $s = 1$ et $r > 1$. Ensuite, une matrice orthogonale Q de taille $p \times p$ est générée de manière uniforme. Les données sont finalement tirées selon une loi normale centrée et de matrice de variance covariance :

$$Q^\intercal \text{diag}(\overbrace{r, \ldots, r}^{d \text{ fois}}, \overbrace{1, ..., 1}^{p-d \text{ fois}})Q,$$

où $p = 50$.

Notre méthodologie (NG, pour normal gamma en anglais) est comparée avec d'autres méthodes d'estimation de d : l'approche de Minka, 2000 basée sur une approximation de Laplace, la validation croisée (GCV) de Josse et Husson, 2012, la technique de profils de vraisemblance (PL) de Zhu et Ghodsi, 2006, et l'algorithme d'optimisation (ML) de Bouveyron, Celeux et Girard, 2011 conçu pour le modèle ACPP contraint utilisé dans

les simulations. Ainsi, pour chaque valeur de n dans $\{40, 50, 70, 100\}$, et une grille de 50 valeurs de SNR entre 1.5 et 30, 50 jeux de données sont simulés et toutes les approches sont appliquées pour estimer d. Les performances sont évaluées à partir du pourcentage du nombre de fois où d est correctement estimé. Comme illustré dans la figure 5.5, les méthodes utilisées pour comparaison ont toutes des résultats qui se dégradent dans au moins un scénario. Par exemple, alors que Laplace permet d'obtenir un estimateur pertinent de d pour $n = 100$ et $n = 70$, elle n'estime jamais correctement d lorsque $n = 40$, pour une plage importante de SNR. A l'inverse, NG a de bonnes performances et des résultats stables, dans tous les scénarios.

Conclusion

Dans ce chapitre, nous nous sommes intéressés à trois problèmes courants dans le cadre de la grande dimension. Le principe est systématiquement le même. Des vecteurs binaires sont tout d'abord intégrés dans des modèles connus afin de caractériser les aspects parcimonieux. Des lois a priori sont alors choisies permettant d'obtenir une forme analytique de la vraisemblance marginale des données. Un chemin de modèles est ensuite construit à l'aide d'algorithmes de type EM simple ou intégrant des aspects variationnels. Enfin, la vraisemblance marginale est maximisée sur le chemin en ajoutant les variables au fur et à mesure. Le premier travail décrit est dédié à la régression linéaire parcimonieuse. Le second autorise la projection de données en ACP à l'aide de vecteurs dont le support est globalement parcimonieux. Le dernier, permet d'estimer le nombre d'axes à utiliser en ACP pour la projection.

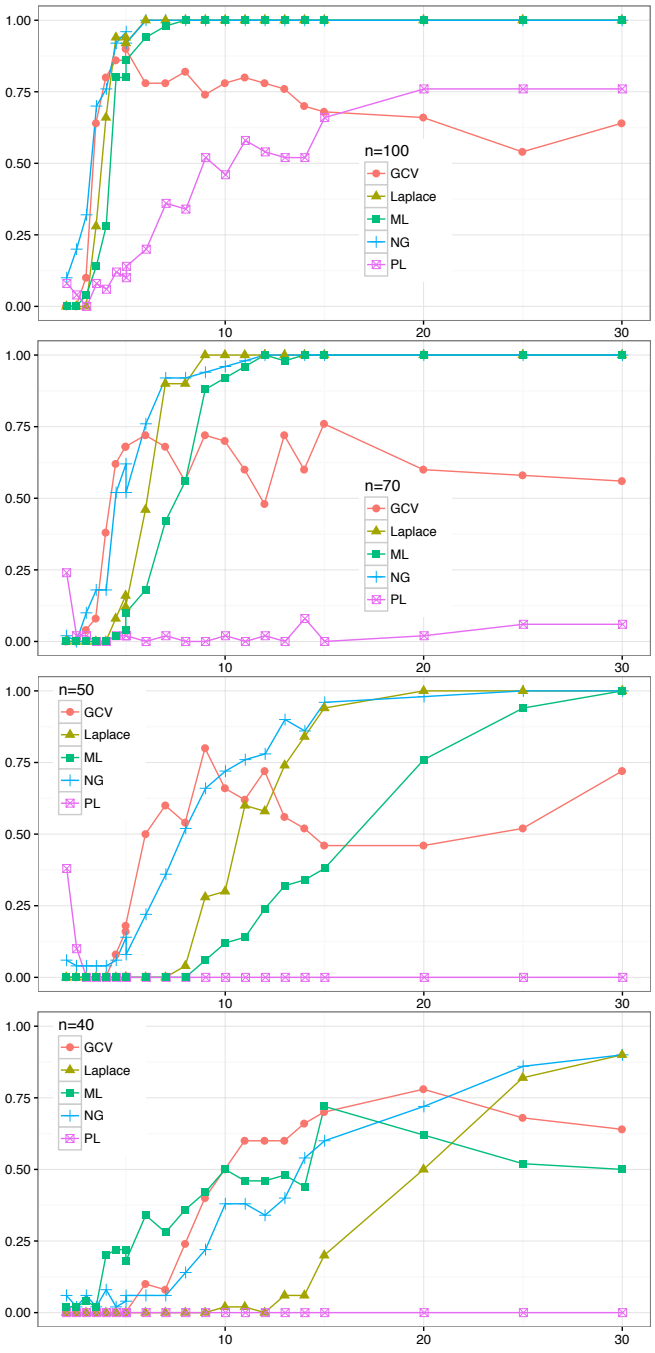


Figure 5.5 – Pourcentage du nombre de fois où la dimension d est correctement estimée, sur 50 répétitions, en fonction du rapport signal sur bruit. De haut en bas : $n = 100,\ 70,\ 50$ et 40.

6

Conclusion et perspectives

Les chapitres précédents reflètent l'état actuel de mes travaux publiés ou, pour quelques uns, sur le point de l'être. Depuis mon affection au laboratoire SAMM de l'université Paris 1 Panthéon-Sorbonne, en septembre 2011, mes recherches ont bien sûr soulevé une série de questions pour des développements futurs. Certains de ces points ont été adressés par moi ou par d'autres au fur et à mesure. Je liste ci-dessous des questions restantes en suspens pour lesquelles je souhaiterais apporter des éléments de solution dans les années à venir.

MODÉLISATION DE RÉSEAUX CREUX Historiquement, la recherche en statistique pour l'analyse des réseaux s'est largement appuyée sur des modèles échangeables (voir discussion dans la section 4.1), comme le modèle SBM ou de W-graphe, généralisant la plupart des autres modèles existants. Malheureusement, la notion d'échangeabilité induit que les graphes générés sont denses ou vides. Or, les réseaux étudiés sont généralement creux, c'est-à-dire que le nombre de connexions observées est faible au regard du nombre de connexions possibles. Les travaux récents et fondateurs de CARON et FOX, à paraître autorisent eux la modélisation de réseaux creux. Le principe consiste à passer d'une notion d'échangeabilité simple à une notion d'échangeabilité sur des mesures aléatoires. Alors qu'une multitude de travaux, souvent dérivés du modèle SBM, ont été proposés dans la littérature afin d'analyser des réseaux de différents types (avec ou sans covariables, connexions binaires, valuées, statiques, dynamiques, ...), relativement peu de travaux ont été publiés à partir de CARON et FOX, à paraître. Mon idée est ici de croiser mes recherches sur les dérivés de SBM et ce modèle pour graphe creux afin de proposer de nouveaux outils d'inférence.

ANALYSE DE DONNÉES DYNAMIQUES DE TYPE RÉSEAU ET TEXTE La plupart des réseaux sociaux décrivent des échanges entre des individus, échanges se faisant par l'intermédiaire de textes. C'est en particulier le cas des réseaux de type Facebook ou Twitter, mais également des réseaux d'emails. Comme indiqué en section 2.5, de manière tout à fait surprenante, les outils d'analyse de réseaux, qui ont pourtant été souvent motivés par des questions pratiques en

sciences sociales, modélisent pour la plupart ces réseaux de manière binaire. Autrement dit, l'information de texte échangé disparaît et seules les présences ou absences de connexions entre individus sont retenues comme source de données. Nous avons donc proposé le modèle STBM (Bouveyron, Latouche et Zreik, 2016) et une procédure d'inférence associée permettant l'analyse conjointe d'un réseau et d'un ensemble de textes portés par les connexions. Malheureusement, ce travail a été réalisé dans un cadre uniquement statique. Or, comme discuté longuement dans le chapitre 3, les échanges sont généralement stockés sous la forme d'interactions. Deux individus interagissent à un temps ν précis qui est enregistré. Lorsque cet échange se fait par l'intermédiaire d'un texte, il est alors nécéssaire de modifier le modèle STBM afin d'y intégrer la notion de processus pour aborder les aspects temporels.

Mélange de d'ACP globalement parcimonieuse La grande dimension soulève de très nombreuses questions aussi bien pratiques que théoriques. Lorsque le nombre p de variables augmente par rapport au nombre d'observations, la géométrie de l'espace des observations devient particulièrement complexe et il est alors nécessaire d'adapter les outils statistiques existants. Concrètement, dans des applications sur données réelles où le nombre de variables peut être de l'ordre du million, le praticien cherche avant toute analyse avancée à réduire p. Nos travaux dans Bouveyron, Latouche et Mattei, 2017a offrent selon moi un cadre principal permettant d'aborder cette question. Sans variable cible particulière et sans a priori sur l'étude à réaliser ensuite, la méthodologie permet en effet de déterminer les variables caractéristiques du jeu de données. Comme nous l'avons montré, le modèle d'ACP probabiliste s'accorde bien avec le cadre Bayésien. Pour des lois a priori particulières, une expression pour la log-vraisemblance marginale des données peut ainsi être obtenue. Il est maintenant fondamental d'étendre ces travaux de manière à pouvoir réaliser la sélection de variables au moment de l'apprentissage statistique. A titre d'exemple, les travaux de Bouveyron, Latouche et Mattei, 2017a doivent être étendus à l'aide des modèles de mélange en vue de faire du clustering des données. Le modèle construit serait donc un mélange d'ACP probabiliste globalement parcimonieuse où chaque composante du mélange aurait ses propres variables explicatives.

Inférence Bayésienne et optimisation Contrairement aux points mentionnés au dessus qui sont dans des cadres bien délimités, j'indique ici une thématique de recherche plus générale pour laquelle je souhaite continuer de contribuer. D'un point de vue personnel, mon année de recherche à l'université d'Aston, avant le début de ma thèse en France, m'a beaucoup marqué et influencé. Dans les années 90, cette université anglaise a en effet été au coeur de travaux précurseurs en statistique, dans un cadre Bayésien, autour de Christopher Bishop, Ian Nabney, et Mike Tipping pour ne citer que quelques noms. Le cadre Bayésien se retrouve ainsi dans presque tous mes travaux. J'ai surtout été profondément marqué par les travaux de David MacKay, décédé l'année dernière. Sa réécriture propre de la régression ridge dans un cadre Bayésien (MacKay, 1992) constitue selon moi un travail absolument fondamental. Une fois la procédure réécrite, les outils d'inférence Bayésienne s'appliquent et peuvent aboutir à des solutions pour le problème considéré, mais également pour des problèmes annexes. Typiquement, l'algorithme d'inférence de MacKay, 1992 autorise l'estimation du paramètre de pénalité de la norme L_2 en même temps que l'estimation du vecteur de régression lui même. Cette philosophie de réécrire dans un cadre Bayésien un problème d'optimisation établi a constitué le point de départ de mes travaux en grande dimension, pour la régression linéaire parcimonieuse et présentés en section 5.2. Nous avons en particulier montré le lien entre la log-vraisemblance marginale obtenue et la fonction objectif utilisée dans l'elastic net de Zou et Hastie, 2005. Ces résultats ont motivé la thèse de

Pierre-Alexandre Mattei et ont produit deux résultats en ACP probabiliste. Je souhaiterais maintenant m'intéresser aux liens entre le cadre Bayésien et la construction de fonctions sous-modulaires pour lesquelles des algorithmes exactes d'optimisation existent. L'idée serait de caractériser une classe de modèles / problèmes pour lesquels la log-vraisemblance marginale des données serait construite à partir de fonctions sous-modulaires.

Bibliographie

K. Pearson. « On lines and planes of closest fit to systems of point in space ». In : *Philosophical Magazine* 2.11 (1901), p. 559–572.

B. de Finetti. « Funzione caratteristica di un fenomeno aletorio. Atti della R ». In : *Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale* 4 (1931), p. 251–299.

H. Hotelling. « Analysis of a complex of statistical variables into principal components. » In : *Journal of educational psychology* 24.6 (1933), p. 417.

H. Jeffreys. « An invariant form for the prior probability in estimations problems ». In : *Proceedings of the Royal Society of London. Series A*. T. 186. 1946, p. 453–461.

E. Hewitt et L.J. Savage. « Symmetric measures on Cartesian products ». In : *Transactions of the American Mathematical Society* 80.2 (1955), p. 470–501.

P. Erdös et A. Rényi. « On random graphs I ». In : *Publicationes Mathematicae Debrecen* 6 (1959), p. 290–297.

R.E. Bellman. *Adaptive control processes : a guided tour*. Princeton university press, 1961.

W. Hoeffding. *The strong law of large numbers for u-statistics*. Rapp. tech. North Carolina State University. Dept. of Statistics, 1961.

M. Abramowitz et I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

H.E. Rauch, F. Tung et T. Striebel. « Maximum likelihood estimates of linear dynamic systems ». In : *AIASS Journal* 3.8 (1965), p. 1445–1450.

R.B. Cattell. « The scree test for the number of factors ». In : *Multivariate behavioral research* 1.2 (1966), p. 245–276.

P. Billingsley. *Convergence of probability measures*. 1968.

H. Akaike. « A new look at the statistical model identification ». In : *IEEE Transactions on Automatic Control* 19.6 (1974), p. 716–723.

J.B. Kruskal. « More factors than subjects, tests and treatments : an indeterminacy theorem for canonical decomposition and individual differences scaling ». In : *Psychometrika* 41.3 (1976), p. 281–293.

A. Dempster, N. Laird et D. Rubin. « Maximum likelihood from incomplete data via the EM algorithm ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), p. 1–38.

J.B. Kruskal. « Three-way arrays : rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics ». In : *Linear algebra and its Applications* 18.2 (1977), p. 95–138.

G. Schwarz. « Estimating the dimension of a model ». In : *The Annals of Statistics* 6.2 (1978), p. 461–464.

S. Wold. « Cross-validatory estimation of the number of components in factor and principal components models ». In : *Technometrics* 20.4 (1978), p. 397–405.

D.N. Hoover. « Relations on probability spaces and arrays of random variables ». In : *Preprint, Institute for Advanced Study, Princeton, NJ* 2 (1979).

D.J. Aldous. « Representations for partially exchangeable arrays of random variables ». In : *Journal of Multivariate Analysis* 11.4 (1981), p. 581–598.

D.W. Scott et J.R. Thompson. « Probability density estimation in higher dimensions ». In : *Computer Science and Statistics : Proceedings of the fifteenth symposium on the interface*. T. 528. North-Holland, Amsterdam. 1983, p. 173–179.

L. Hubert et P. Arabie. « Comparing partitions ». In : *Journal of classification* 2.1 (1985), p. 193–218.

Y.J. Wang et G.Y. Wong. « Stochastic blockmodels for directed graphs ». In : *Journal of the American Statistical Association* 82 (1987), p. 8–19.

T.J. Mitchell et J.J. Beauchamp. « Bayesian variable selection in linear regression (with discussion) ». In : *Journal of the American Statistical Association* 83 (1988), p. 1023–1036.

M.A. Arcones et E. Gine. « On the bootstrap of U and V statistics ». In : *The Annals of Statistics* (1992), p. 655–674.

E. Gine et J. Zinn. « Marcinkiewicz type laws of large numbers and convergence of moments for U-statistics ». In : *Probability in Banach Spaces, 8 : Proceedings of the Eighth International Conference*. Springer. 1992, p. 273–291.

D.J.C. MacKay. « Bayesian interpolation ». In : *Neural computation* 4.3 (1992), p. 415–447.

R.E. Kass et A.E. Raftery. « Bayes factors ». In : *Journal of the american statistical association* 90.430 (1995), p. 773–795.

R. Tibshirani. « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* 58.1 (1996), p. 267–288.

S. Roweis. « EM algorithms for PCA and SPCA ». In : *Advances in neural information processing systems* (1998), p. 626–632.

J.A. Hoeting, D. Madigan, A.E. Raftery et C.T. Volinsky. « Bayesian model averaging : a tutorial ». In : *Statistical science* (1999), p. 382–401.

M.E. Tipping et C.M. Bishop. « Probabilistic principal component analysis ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 61.3 (1999), p. 611–622.

C. Biernacki, G. Celeux et G. Govaert. « Assessing a mixture model for clustering with the integrated completed likelihood ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7 (2000), p. 719–725.

T.S. Jaakkola et M.I. Jordan. « Bayesian parameter estimation via variational methods ». In : *Statistics and Computing* 10 (2000), p. 25–37.

T.P. Minka. « Automatic choice of dimensionality for PCA ». In : *Advances in Neural Information Processing Systems*. T. 13. 2000, p. 598–604.

J. Shi et J. Malik. « Normalized cuts and image segmentation ». In : *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), p. 888–905.

A.W. Van der Vaart. *Asymptotic statistics*. T. 3. Cambridge university press, 2000.

K. Nowicki et T.A.B. Snijders. « Estimation and prediction for stochastic blockstructures ». In : *Journal of the American Statistical Association* 96.455 (2001), p. 1077–1087.

P.D. Hoff, A.E. Raftery et M.S. Handcock. « Latent space approaches to social network analysis ». In : *Journal of the American Statistical Association* 97.460 (2002), p. 1090–1098.

P.D. Hoff, A.E. Raftery et M.S. Handcock. « Latent space approaches to social network analysis ». In : *Journal of the american Statistical association* 97.460 (2002), p. 1090–1098.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

D.M. Blei, A.Y. Ng et M.I. Jordan. « Latent dirichlet allocation ». In : *Journal of machine Learning research* 3 (2003), p. 993–1022.

G. McLachlan et D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

H. Zou et T. Hastie. « Regularization and variable selection via the elastic net ». In : *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* 67.2 (2005), p. 301–320.

L. László et B. Szegedy. « Limits of dense graph sequences ». In : *Journal of Combinatorial Theory, Series B* 96.6 (2006), p. 933–957.

M.E.J. Newman. « Modularity and community structure in networks ». In : *Proceedings of the national academy of sciences* 103.23 (2006), p. 8577–8582.

G. Palla, I. Derenyi, I. Farkas et T. Vicsek. *CFinder the community/cluster finding program*. Version 2.0.1. 2006.

M. Zhu et A. Ghodsi. « Automatic dimensionality selection from the scree plot via the use of profile likelihood ». In : *Computational Statistics & Data Analysis* 51.2 (2006), p. 918–930.

H. Zou. « The adaptive lasso and its oracle properties ». In : *Journal of the American Statistical Association* 101.476 (2006), p. 1418–1429.

H. Zou, T. Hastie et R. Tibshirani. « Sparse principal component analysis ». In : *Journal of computational and graphical statistics* 15.2 (2006), p. 265–286.

C. Bishop. « Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn ». In : *Springer, New York* (2007).

K. Heller et Z. Ghahramani. « A nonparametric Bayesian approach to modeling overlapping clusters ». In : *Proceedings of the 11th International Conference on AI and Statistics*. 2007.

P.D. Hoff. « Model averaging and dimension selection for the singular value decomposition ». In : *Journal of the American Statistical Association* 102.478 (2007), p. 674–685.

E.M. Airoldi, D.M. Blei, S.E. Fienberg et E.P. Xing. « Mixed membership stochastic blockmodels ». In : *Journal of Machine Learning Research* 9 (2008), p. 1981–2014.

F. Caron et A. Doucet. « Sparse Bayesian nonparametric regression ». In : *Proceedings of the 25th International Conference on Machine Learning*. 2008.

J-J. Daudin, F. Picard et S. Robin. « A mixture model for random graphs ». In : *Statistics and Computing* 18.2 (2008), p. 173–183.

P. Diaconis et S. Janson. « Graph limits and exchangeable random graphs ». In : *Rendiconti di Matematica e delle sue Applicazioni* 7.28 (2008), p. 33–61.

K. Heller, S. Williamson et Z. Ghahramani. « Statistical models for partial membership ». In : *Proceedings of the 25th International Conference on Machine Learning*. 2008, p. 392–399.

J.M. Hofman et C.H. Wiggins. « Bayesian approach to network modularity ». In : *Physical review letters* 100.25 (2008), p. 258701.

E.S. Allman, C. Matias et J.A. Rhodes. « Identifiability of parameters in latent structure models with many observed variables ». In : *Annals of Statistics* 37.6A (2009), p. 3099–3132.

C. Archambeau et F. Bach. « Sparse probabilistic projections ». In : *Advances in neural information processing systems*. 2009, p. 73–80.

R. Jenatton, G. Obozinski et F. Bach. « Structured sparse principal component analysis ». In : *International Conference on Artificial Intelligence and Statistics*. 2009.

P. Latouche, E. Birmelé et C. Ambroise. « Bayesian methods for graph clustering ». In : *Advances in Data Handling and Business Intelligence*. Springer, 2009.

A. Goldenberg, A.X. Zheng, S.E. Fienberg, E.M. Airoldi et al. « A survey of statistical network models ». In : *Foundations and Trends® in Machine Learning* 2.2 (2010), p. 129–233.

A.A. Gouda et T. Szántai. « On numerical calculation of probabilities according to Dirichlet distribution ». In : *Annals of Operations Research* 177 (2010). DOI : 10.1007/s10479-009-0601-9, p. 185–200.

M. Mariadassou, S. Robin et C. Vacher. « Uncovering latent structure in valued graphs : a variational approach ». In : *Annals of Applied Statistics* 4.2 (2010), p. 715–742.

N. Meinshausen et P. Bühlmann. « Stability selection ». In : *Journal of the Royal Statistical Society : Series B* 27 (2010).

N.X. Vinh, J. Epps et J. Bailey. « Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance ». In : *Journal of Machine Learning Research* 11.Oct (2010), p. 2837–2854.

P.J. Bickel, A. Chen et E. Levina. « The method of moments and degree distributions for network models ». In : *Annals of Statistics* 39.5 (2011), p. 2280–2301.

C. Bouveyron, G. Celeux et S. Girard. « Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA ». In : *Pattern Recognition Letters* 32.14 (2011), p. 1706–1713.

Q. Ho, L. Song et E.P. Xing. « Evolving cluster mixed-membership blockmodel for time-evolving networks ». In : *International Conference on Artificial Intelligence and Statistics*. 2011, p. 342–350.

P. Latouche, E. Birmelé et C. Ambroise. « Overlapping stochastic block models with application to the French political blogosphere ». In : *Annals of Applied Statistics* 5.1 (2011), p. 309–336.

F. Bach, R. Jenatton, J. Mairal et G. Obozinski. « Optimization with sparsity-inducing penalties ». In : *Foundations and Trends in Machine Learning* 4.1 (2012), p. 1–106.

J. Josse et F. Husson. « Selecting the number of components in principal component analysis using cross-validation approximations ». In : *Computational Statistics & Data Analysis* 56.6 (2012), p. 1869–1879.

R. Killick, P. Fearnhead et I. A. Eckley. « Optimal Detection of Changepoints With a Linear Computational Cost ». In : *Journal of the American Statistical Association* 107.500 (2012), p. 1590–1598. doi : 10.1080/01621459.2012.737745. eprint : http://dx.doi.org/10.1080/01621459.2012.737745. url : http://dx.doi.org/10.1080/01621459.2012.737745.

P. Latouche, E. Birmele et C. Ambroise. « Variational Bayesian inference and complexity control for stochastic block models ». In : *Statistical Modelling* 12.1 (2012), p. 93–115.

P.K. Gopalan et D.M. Blei. « Efficient discovery of overlapping communities in massive networks ». In : *Proceedings of the National Academy of Sciences* 110.36 (2013), p. 14534–14539.

D. Hernández-Lobato, J.M. Hernández-Lobato et P. Dupont. « Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation ». In : *Journal of Machine Learning Research* 14.1 (2013), p. 1891–1945.

A.F. McDaid, T.B. Murphy, N. Friel et N.J. Hurley. « Improved Bayesian inference for the stochastic block model with application to large networks ». In : *Computational Statistics & Data Analysis* 60 (2013), p. 12–31.

L. Nouedoui et P. Latouche. « Bayesian non parametric inference of discrete valued networks ». In : *21-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*. Bruges, Belgium, 2013, p. 291–296. url : https://hal.archives-ouvertes.fr/hal-00825966.

M. Gavish et D.L. Donoho. « The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$ ». In : *IEEE Transactions on Information Theory* 60.8 (2014), p. 5040–5053.

Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou et S. Lamassé. « The random subgraph model for the analysis of an ecclesiastical network in Merovingian Gaul ». In : *Annals of Applied Statistics* 8.1 (2014), p. 377–405.

P. Latouche, E. Birmelé et C. Ambroise. « Model selection in overlapping stochastic block models ». In : *Electronic Journal of Statistics* 8.1 (2014), p. 762–794.

P. Latouche, E. Birmelé et C. Ambroise. « Overlapping clustering methods for networks ». In : *Handbook of Mixed Membership Models and Their Applications*. Chapman et Hall/CRC, 2014.

C. Matias et S. Robin. « Modeling heterogeneity in random graphs through latent space models : a selective review ». In : *ESAIM : Proceedings and Surveys* 47 (2014), p. 55–74.

E. Côme et P. Latouche. « Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood ». In : *Statistical Modelling* 15.6 (2015), p. 564–589.

R. Guigourès, M. Boullé et F. Rossi. « Discovering patterns in time-varying graphs : a triclustering approach ». English. In : *Advances in Data Analysis and Classification* (2015), p. 1–28. issn : 1862-5347. doi : 10.1007/s11634-015-0218-6. url : http://dx.doi.org/10.1007/s11634-015-0218-6.

T. Hastie, R. Tibshirani et M. Wainwright. *Statistical Learning with Sparsity : The Lasso and Generalizations*. CRC Press, 2015.

C. Matias, T. Rebafka et F. Villers. « A semiparametric extension of the stochastic block model for longitudinal networks ». In : *arXiv preprint arXiv :1512.07075* (2015).

R. Zreik, P. Latouche et C. Bouveyron. « Classification automatique de réseaux dynamiques avec sous-graphes : étude du scandale Enron ». In : *Journal de la Société Française de Statistique* 156.3 (2015), p. 166–191.

R. Zreik, P. Latouche et C. Bouveyron. « Cluster Identification in Maritime Flows with Stochastic Methods ». In : *Maritime Networks : Spatial Structures and Time Dynamics*. Routledge, 2015.

C. Bouveyron, P. Latouche et R. Zreik. « The stochastic topic block model for the clustering of vertices in networks with textual edges ». In : *Statistics and Computing* (2016), p. 1–21.

M. Corneli, P. Latouche et F. Rossi. « Block modelling in dynamic networks with non-homogeneous Poisson processes and exact ICL ». In : *Social Network Analysis and Mining* 6.1 (2016), p. 55–85.

M. Corneli, P. Latouche et F. Rossi. « Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks ». In : *Neurocomputing* 192 (2016), p. 81–91.

P. Latouche, P-A. Mattei, C. Bouveyron et J. Chiquet. « Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression ». In : *Journal of Multivariate Analysis* 146 (2016), p. 177–190.

P. Latouche et S. Robin. « Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models ». In : *Statistics and Computing* 26.6 (2016), p. 1173–1185.

L. Yengo, J. Jacques, C. Biernacki et M. Canouil. « Variable clustering in high-dimensional linear regression : The r package clere ». In : *R Journal* 8.1 (2016), p. 92–106.

R. Zreik, P. Latouche et C. Bouveyron. « The dynamic random subgraph model for the clustering of evolving networks ». In : *Computational Statistics* (2016), p. 1–33.

C. Bouveyron, P. Latouche et P-A. Mattei. « Bayesian Variable Selection for Globally Sparse Probabilistic PCA ». 2017.

C. Bouveyron, P. Latouche et P-A. Mattei. « Exact dimensionality selection for Bayesian PCA ». 2017.

S. Ouadah, S. Robin et P. Latouche. « A degree-based goodness-of-fit test for heterogeneous random graph models ». 2017.

J. Wyse, N. Friel et P. Latouche. « Inferring structure in bipartite networks using the latent blockmodel and exact ICL ». In : *Network Science* 5.1 (2017), p. 45–69.

F. Caron et E.B. Fox. « Sparse graphs using exchangeable random measures ». In : *Journal of the Royal Statistical Society* (à paraître).

M. Corneli, P. Latouche et F. Rossi. « Multiple change point detection and clustering in dynamic networks ». In : *Statistics and Computing* (à paraître).

P. Latouche, S. Robin et S. Ouadah. « Goodness of fit of logistic regression models for random graphs ». In : *Journal of Computational and Graphical Statistics* (à paraître).

C. Matias et V. Miele. « Statistical clustering of temporal networks through a dynamic stochastic block model ». In : *Journal of the Royal Statistical Society : Series B* (à paraître).