Inférence de réseaux pour modèles inflatés en zéro

Clémence Karmann

▶ **To cite this version:**

Clémence Karmann. Inférence de réseaux pour modèles inflatés en zéro. Statistiques [math.ST]. Université de Lorraine, 2019. Français. NNT : 2019LORR0146 . tel-02384511

HAL Id: tel-02384511
https://hal.science/tel-02384511

Submitted on 28 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Inférence de réseaux pour modèles inflatés en zéro

THÈSE

présentée et soutenue publiquement le 25 novembre 2019

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention mathématiques)

par

Clémence Karmann

Composition du jury

Président :	Pascal Moyal	Université de Lorraine
Directrices :	Anne Gégout-Petit	Université de Lorraine
	Aurélie Gueudin-Muller	Université de Lorraine
Rapporteurs :	Stéphane Chrétien	Université de Lyon 2
	Marie-Laure Martin-Magniette	INRA MIA Paris - AgroParisTech
Examinateurs :	Julien Chiquet	INRA MIA Paris - AgroParisTech
	Fanny Villers	LPSM, Sorbonne Université

Institut Elie Cartan de Lorraine
Institut National de Recherche en Informatique et en Automatique

Mis en page avec la classe thesul.

Remerciements

Ces remerciements sont loin d'être la partie la plus facile à écrire et je tiens à présenter toutes mes excuses aux personnes que je vais oublier de mentionner, et qui ont pourtant rendu cette période plus douce.

Tout d'abord, je souhaite remercier chaleureusement Anne et Aurélie, qui ont accepté de vivre cette aventure avec moi, et grâce à qui j'ai énormément appris sur le plan mathématique, comme sur le plan humain. J'ai eu beaucoup de chance de travailler avec elles et j'ai grandement apprécié leur humanité, leur bienveillance, leur éternel optimisme, ainsi que leurs remarquables écoute et soutien qui ont été cruciaux dans les moments difficiles. La confiance qu'elles m'ont accordée tout au long de ces trois années me touche profondément.

Je tiens également à remercier Stéphane Chrétien et Marie-Laure Martin-Magniette, qui ont accepté de rapporter ma thèse, ainsi que tous les autres membres du jury, Julien Chiquet, Pascal Moyal et Fanny Villers. Je suis très honorée qu'ils aient accepté de venir donner leurs avis d'expert sur mon travail, particulièrement en cette fin de mois de novembre dans la grisaille de l'Est.

Merci aux professeurs qui m'ont appris à faire des mathématiques et m'ont donné envie (parfois le courage) de poursuivre sur cette voie, la liste n'étant pas exhaustive : Marie-Line Chabanol, Pierre-Emmanuel Chaput, François Chargois, Philippe Chassaing, Benoît Daniel, David Dos Santos, Bruno Duchesne, Lucas Fresse, Olivier Garet, Françoise Geandier, Alain Genestier, Jean-Sébastien Giet, Jean-Francois Grosjean, Oussama Hijazi, Khalid Koufany, Vincent Koziarz, Aline Kurtzmann, Manfred Madritsch, Julien Maubon, Séraphin Méfire, Frédéric Robert, Thomas Stoll, Gérald Tenenbaum, Matei Toma. Petite mention spéciale à Régine Marchand pour ses encouragements permanents, à Didier Schmitt pour m'avoir pris sous son aile, à Michel Matignon dont le soutien et l'investissement m'ont (et continuent) de me toucher infiniment. Il s'avère que j'ai eu la chance de continuer à côtoyer beaucoup d'entre eux encore pendant ma thèse, et j'ai été émue par leurs réguliers encouragements et marques de sympathie. Sur le même plan, j'aimerais également remercier toutes les personnes qui rendent ce laboratoire particulièrement accueillant et nos conditions de travail exceptionnelles, notamment Nathalie Benito, Elodie Cunat, Laurence Quirot et Paola Schneider, toujours prêtes à se plier en quatre pour nous aider, Didier Gemmerlé, dont j'ai trop souvent confondu le bureau avec le mien, Céline Cordier, les chaleureux membres de l'équipe de probabilités et statistiques et de BIGS. Merci également à Bruno Pinson et Antoine Henrot, qui n'ont pas hésité à me donner un coup de main sur des problèmes d'optimisation, à Jean-Marc Sac-Epée, qui m'a toujours dépannée avec le babycluster, aux enseignants-chercheurs avec qui j'ai eu le privilège de travailler, notamment Irène Marcovici et Pascal Moyal.

Merci également à toutes les personnes avec qui j'ai pu échanger en recherche et qui ont, avec beaucoup de bienveillance, aidé à améliorer ce travail et suggéré de nombreuses pistes : Julien Chiquet, Stéphane Chrétien, Emilie Devijver, Stéphane Robin, Nathalie Villa... Merci à Guillaume Hureaux pour son investissement et sa bonne humeur.

Un merci particulier pour la gentillesse et l'immense patience de Takéo, pour la rigueur, le dévouement, l'amitié, les explications et analyses aussi éclairantes que pertinentes de Romain Azaïs, l'oreille bienveillante de Vladimir Latocha, pour la "présence" et le soutien de mes co-

bureaux, Florian puis Robin, pour les idées lumineuses de Pierre-Adrien, pour les blagues et histoires loufoques d'Éloïse et Coralie, pour le soutien et la bonne humeur des doctorants et post-doctorants dont j'ai croisé la route, notamment Pierre, Matthieu, Dimitry, Johann, Rodolphe, Ulysse, Vincent, Philippe, Paolo, Thomas, Gianluigi, Nassim, Christophe...

Sur un plan plus personnel, je souhaite remercier ma famille (pour tellement de choses), en particulier mes parents, qui m'ont transmis la fibre des maths (ahah) et qui supportent mes pires défauts, et ma sœur, pour les leçons de vie qu'elle me donne sans s'en rendre compte. Je remercie tous mes amis qui m'ont régulièrement apporté du réconfort : Alexandra, Beubeu, Camille, Dorine, Eliane, Florine, Frédéric, JB, Karen, Magali, Pierre, Pierrot, Sarah. Merci puissance mille à Valentin, Elise, Julien, Margot et Hadrien pour leur soutien sans limites et leur amour inconditionnel.

Je ne pourrais conclure ces remerciements sans mentionner Tom, pour tout ce qu'il m'a apporté, pour son infaillible soutien, sa confiance permanente, sa désinvolture, pour son intuition mathématique et son incroyable capacité de réflexion qui ne cesseront jamais de m'impressionner. Ce manuscrit n'aurait jamais existé sans lui.

Brève introduction générale

Les travaux de ce manuscrit sont liés à l'inférence de réseaux pour des données "non-standard" dans le sens où elles sont inflatées en zéro, c'est-à-dire qu'elles comportent un fort taux de zéros comme c'est par exemple le cas des données d'abondance ou de comptage. L'inférence de réseaux ou inférence de graphes a de plus en plus d'applications notamment en santé humaine et en environnement pour l'étude de données micro-biologiques et génomiques. Les réseaux constituent en effet un outil approprié pour représenter, voire étudier des relations entre des entités.

Ces travaux ont en fait été initialement motivés par une collaboration entre l'équipe BIGS d'INRIA Nancy et Aurélie Deveau de l'INRA Nancy sur des données d'abondance de micro-organismes, inflatées en zéros. Le traitement des données d'abondance est particulier pour deux raisons : d'une part elles ne reflètent pas directement la réalité car un processus de séquençage a lieu pour dupliquer les espèces et ce processus apporte de la variabilité, d'autre part une espèce peut être absente dans certains échantillons. On est alors dans le cadre de données inflatées en zéro. Ces particularités rendent la modélisation de l'inflation en zéro assez complexe et les modèles inflatés en zéro restent encore très peu étudiés alors qu'ils reflètent la structure de nombreux jeux de données de façon pertinente.

Par ailleurs, en inférence de réseaux, beaucoup de méthodes ont été développées sur des modèles présentant des propriétés intéressantes. C'est le cas notamment du modèle graphique gaussien mais aussi du modèle d'Ising pour les données binaires. Ces modèles présentent de belles propriétés théoriques sur l'indépendance conditionnelle, propices à l'inférence de réseaux. Les difficultés liées à la modélisation et l'étude de modèles inflatés en zéro nous ont incités à nous appuyer sur des modèles existants et adaptés à l'inférence de réseaux, comme c'est le cas des modèle graphique gaussien et modèle d'Ising que nous présentons en détails dans une première partie. Les méthodes d'inférence existantes pour ces deux modèles nous ont amenés à considérer le cas de régressions ordinales pénalisées et à développer une méthode de sélection de variables. Ces outils seront développés dans une seconde partie, l'objectif étant d'inférer un réseau en estimant les voisinages par une procédure couplant des méthodes de régressions ordinales et de sélection de variables. La troisième et dernière partie se focalise sur l'inférence de réseaux dans un modèle où les variables sont des gaussiennes inflatées en zéro par double troncature (à droite et à gauche).

Table des matières

Brève introduction générale **iii**

Partie I État de l'art **1**

Chapitre 1
Modèle graphique gaussien

1.1 Définitions et propriétés 6
1.1.1 Rappels sur le modèle de régression linéaire gaussien 6
1.1.2 Généralités 7
1.1.3 Lien avec la matrice de précision 8
1.1.4 Lien avec les coefficients des régressions linéaires 9
1.1.5 Lien avec les corrélations partielles 10
1.2 Approches statistiques pour l'inférence de réseaux 14
1.2.1 Méthodes basées sur l'estimation des matrices de covariance ou précision 15
1.2.2 Méthodes basées sur l'estimation d'une approximation du graphe de concentration 17
1.2.3 Méthodes basées sur l'estimation des voisinages 18

Chapitre 2
Modèles binaires

2.1 Régression logistique 21
2.1.1 Généralités 21
2.1.2 Modèle linéaire généralisé et pénalisations 22
2.2 Le modèle d'Ising 25
2.2.1 Brève présentation du modèle d'Ising 25
2.2.2 Lien avec la régression logistique 26
2.2.3 Simulation de données suivant la loi d'Ising 27
2.2.4 Extensions du modèle d'Ising 29

2.3 Approches statistiques 31
2.3.1 Sur la régression logistique pénalisée 31
2.3.2 Dans le cas de l'inférence de réseaux 31
2.3.3 Comparaisons pratiques des méthodes Glasso et régression logistique 32

Principales contributions **35**

Partie II Régression pénalisée à logits cumulatifs et sélection de variables par la méthode des knockoffs **39**

Chapitre 3
Régression L_1-pénalisée à logits cumulatifs et odds proportionnels

3.1 Modèle à logits cumulatifs (et odds proportionnels) 43
3.1.1 Généralités 43
3.1.2 Interprétation des coefficients 46
3.2 Estimation et inférence 46
3.2.1 Estimation Lasso des coefficients $\boldsymbol{\beta}$ 47
3.2.2 Paramètre de pénalisation et sélection de variables 48
3.3 Simulations 52
3.3.1 Validation croisée 52
3.3.2 Distributions des covariables 53
3.3.3 Stability selection 54
3.3.4 Knockoffs revisités 58
3.4 Conclusions 65

Chapitre 4
La méthode des knockoffs revisités pour la sélection de variables

4.1 Méthode des knockoffs revisités 67
4.1.1 Contexte 67
4.1.2 Principe et généralités 69
4.1.3 Choix du seuil 70
4.1.4 Package R `kosel` 72
4.2 Simulations 73
4.2.1 Paramètres de simulations 73
4.2.2 Efficacité et comparaisons - $p = 50$ 74
4.2.3 Efficacité et comparaisons - $p = 2000$ 78
4.2.4 Caractère aléatoire de la procédure 82

4.3 Conclusions 84

Chapitre 5
Application de modèles de régression pour variables ordinales et de la méthode des knockoffs revisités à l'inférence de réseaux pour données inflatées en zéro

5.1 Modèles pour la simulation de données inflatées en zéro 86
5.1.1 Données gaussiennes inflatées en zéro par des Bernoullis 87
5.1.2 Données *adjacent* 87
5.2 Méthode d'inférence 89
5.3 Résultats 90
5.3.1 Données gaussiennes inflatées en zéro par des Bernoullis 90
5.3.2 Données *adjacent* 92
5.4 Comparaisons avec d'autres méthodes 92
5.4.1 Données gaussiennes inflatées en zéro par des Bernoullis 94
5.4.2 Données *adjacent* 95
5.5 Application à des données réelles 97
5.6 Conclusions 99

Partie III Modèles inflatés en zéro 101

Chapitre 6
Inférence de réseaux pour données gaussiennes inflatées en zéro par double troncature

6.1 Modèle théorique et procédure d'estimation 104
6.1.1 Présentation du modèle 104
6.1.2 Quelques outils théoriques 105
6.1.3 Estimation 106
6.2 Résultats de convergence 109
6.2.1 Estimateurs des points de troncature 109
6.2.2 Estimateur de la matrice de covariance 111
6.2.3 Estimateur de la matrice de précision 116
6.3 Simulations 119
6.3.1 Paramètres de simulations 119
6.3.2 Choix du paramètre de pénalisation 121
6.3.3 Efficacité de la procédure 123

6.3.4 Impact de l'estimation des points de troncature 125
6.3.5 Impact des points de troncature 127
6.3.6 Autres structures de graphes 129
6.4 Conclusions 132

Conclusion générale et perspectives **135**

Annexe A
Preuves et compléments du chapitre 6

Bibliographie **143**

Première partie

État de l'art

Un graphe (V, E) est un objet mathématique composé de deux ensembles : l'ensemble $V = \{v_1, \dots, v_p\}$ des nœuds, qui représentent les entités, et un ensemble E d'arêtes entre ces nœuds, qui modélisent des relations entre ces entités. L'ensemble E est en fait un sous-ensemble des paires de sommets c'est-à-dire $E \subset \{\{v_i, v_j\} : i, j \in \{1, \dots, p\}, i \neq j\}$. Un graphe est parfois aussi appelé réseau par abus de langage. En effet, un réseau est un graphe dont les sommets et/ou arêtes comportent des attributs (par exemple, des noms). Dans la suite, on utilisera indifféremment les deux dénominations.

Les réseaux sont de plus en plus utilisés pour représenter et étudier des relations entre des entités dans des domaines très variés comme la sociologie, la biologie, la santé humaine, l'environnement... En sociologie, on peut faire un réseau d'individus ; dans ce cas, une arête entre deux individus signifie que ceux-ci se connaissent. En biologie, les réseaux de gènes sont devenus très populaires pour mieux étudier et comprendre les interactions entre gènes ; les arêtes représentent alors un lien de dépendance dans l'expression des deux gènes.

Les graphes dont on a parlé jusqu'ici sont les plus basiques mais il existe des types de graphes un peu plus élaborés :

— les graphes orientés : chaque arête a un sens et modélise le plus souvent l'influence d'une variable sur une autre. Dans ce cas, l'ensemble des arêtes est un sous-ensemble des couples de sommets $E \subset \{(v_i, v_j) : i, j \in \{1, \dots, p\}, i \neq j\}$. Dans ce type de graphes, on peut avoir $(v_i, v_j) \in E$ et $(v_j, v_i) \notin E$.

— les graphes pondérés : chaque arête a un poids (positif) qui modélise l'intensité du lien entre les variables.

Sauf mention contraire, les graphes considérés ici seront non orientés et non pondérés.

Ainsi, dans notre cas, le graphe $G = (V, E)$ est constitué d'un ensemble fini de sommets $V = \{v_1, \dots, v_p\}$ et d'arêtes $E \subset \{\{v_i, v_j\} : i, j \in \{1, \dots, p\}, i \neq j\}$. À tout graphe fini G, on peut associer une matrice $A \in \mathcal{M}_p(\mathbb{R})$ telle que $A_{i,i} = 0$ pour tout $i \in \{1, \dots, p\}$ et $A_{i,j} = 1$ si et seulement si $\{v_i, v_j\} \in E$. Cette matrice A est symétrique et est appelée matrice d'adjacence, elle caractérise en fait exactement le graphe. Plus précisément, on a une bijection entre l'ensemble des graphes finis de taille p non orientés et non pondérés et l'ensemble des matrices symétriques de diagonale nulle et d'entrées nulles ou égales à 1.
Cette notion de matrice d'adjacence se généralise aisément aux graphes orientés (la matrice n'est alors plus forcément symétrique) ou aux graphes pondérés (les 1 sont alors remplacés par les poids).

Supposons que l'on connaisse la valeur de p variables quantitatives sur n sujets (dans beaucoup d'applications, $n \ll p$). Les données se présentent sous la forme d'une matrice X comme suit :

$$X = \begin{pmatrix} X_1^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & \dots & X_p^{(2)} \\ & \dots & \\ X_1^{(n)} & \dots & X_p^{(n)} \end{pmatrix},$$

où $X_j^{(i)}$ désigne la i-ème observation de la variable X_j. Les lignes représentent les n observations

et les colonnes les p variables. Les variables sont représentées par les sommets du graphe et une arête entre deux sommets représente un lien (à préciser) entre les variables associées. L'inférence de réseaux consiste alors à déterminer quelles sont les arêtes du réseau. Selon les domaines, l'interprétation de ces arêtes est différente et souvent, les méthodes statistiques relatives à l'inférence également. Nathalie Villa donne une excellente introduction aux réseaux et à leur inférence dans [108]. La plupart du temps, les arêtes modélisent un lien de dépendance et c'est ce type de liens auquel on s'intéresse ici.

Butte et Kohane (1999) [12] et (2000) [11] ont proposé une première approche, basique et naïve, appelée *relevance network*. Elle consiste à calculer les corrélations empiriques entre toutes les paires de variables et à seuiller ensuite pour ne garder que les corrélations les plus fortes, éventuellement à l'aide d'un test. Bien que cette approche soit facile à interpréter, elle est également source de mauvaises interprétations. On s'intéresse essentiellement à représenter les interactions directes ; or dans ce cas, les liens indirects sont aussi détectés. Par exemple, si une variable X_1 influe directement sur deux variables X_2 et X_3 sans que X_2 et X_3 ne soient directement liées (voir figure 1), cette approche va détecter ce lien indirect. En statistique, ce problème est connu sous le nom de paradoxe de Simpson [31], notamment dans le cadre de variables qualitatives. La variable X_1 est alors appelée "facteur de confusion". Si on s'intéresse à des variables quantitatives, le modèle suivant donne un exemple d'un autre type de lien indirect : si $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2, 3$ sont indépendantes, en posant $X_1 = \epsilon_1$, $X_2 = 2X_1 + \epsilon_2$ et $X_3 = X_2 + \epsilon_3$, on obtient alors que $\text{cor}(X_1, X_3) = 2/\sqrt{6}$ tandis que $\text{cor}(X_1, X_3|X_2) = 0$. En terme d'interprétation, ceci

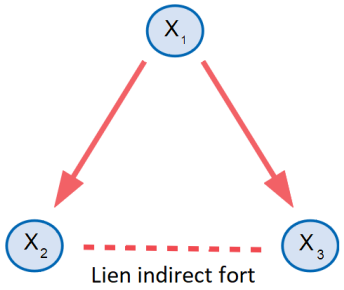


FIGURE 1 – Illustration de la détection de liens indirects.

peut être très problématique et peut conduire à d'importants malentendus et contresens.

Pour pallier ce problème dans le cadre de variables quantitatives, une solution plus pertinente est de s'intéresser aux corrélations partielles c'est-à-dire aux corrélations de chaque paire de variables sachant les autres variables, supposées fixées. Cette notion sera définie plus rigoureusement par la suite. Dans l'idée, cette corrélation partielle correspond à regarder la variation de deux variables en s'affranchissant de la variation due à l'influence des variables restantes. Dans certains cadres, notamment le cadre gaussien développé par la suite, ces corrélations partielles présentent des aspects théoriques remarquables.

Chapitre 1

Modèle graphique gaussien

Sommaire

1.1 Définitions et propriétés 6
1.1.1 Rappels sur le modèle de régression linéaire gaussien 6
1.1.2 Généralités 7
1.1.3 Lien avec la matrice de précision 8
1.1.4 Lien avec les coefficients des régressions linéaires 9
1.1.5 Lien avec les corrélations partielles 10
1.2 Approches statistiques pour l'inférence de réseaux 14
1.2.1 Méthodes basées sur l'estimation des matrices de covariance ou précision 15
1.2.2 Méthodes basées sur l'estimation d'une approximation du graphe de concentration 17
1.2.3 Méthodes basées sur l'estimation des voisinages 18

Contexte :

On dispose de données qu'on présente sous la forme d'une matrice X comme suit :

$$X = \begin{pmatrix} X_1^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & \dots & X_p^{(2)} \\ & \dots & \\ X_1^{(n)} & \dots & X_p^{(n)} \end{pmatrix},$$

où les observations, c'est-à-dire chaque ligne $i \in \{1, \dots, n\}$ de la matrice X, sont supposées indépendantes. Dans le modèle graphique gaussien, décrit notamment par Whittaker (1990) [106], Lauritzen (1996) [56], Edwards (2000) [31] ou encore Hastie, Tibshirani et Friedman (2001) [44], on fait l'hypothèse que chaque observation est issue d'une loi $\mathcal{N}_p(\mu, \Sigma)$:

$$\forall i \in \{1, \dots, n\},\ X^{(i)} = (X_1^{(i)}, \dots, X_p^{(i)}) \sim \mathcal{N}_p(\mu, \Sigma). \tag{1.1}$$

Le but est alors d'inférer un réseau, dit de dépendances conditionnelles ou d'indépendance conditionnelle, entre les variables $X_1, X_2, \dots, X_p$. Ce réseau est défini à la sous-section 1.1.2.

1.1 Définitions et propriétés

1.1.1 Rappels sur le modèle de régression linéaire gaussien

Le modèle de régression linéaire [84, 109] est un outil statistique habituellement mis en œuvre pour l'étude de données multidimensionnelles. Ce modèle comprend une variable aléatoire Y à expliquer et un certain nombre p de variables dites explicatives $X_1, \ldots, X_p$, qui ne sont en général pas considérées comme aléatoires. En pratique, on dispose d'un échantillon de n observations indépendantes de ces variables : $(Y^{(i)}, X_1^{(i)}, \ldots, X_p^{(i)})$ pour $i = 1, \ldots, n$.

Définition 1.1.1 (Régression linéaire multiple)**.** *Soit Y une variable aléatoire quantitative à expliquer par p variables explicatives réelles $X_1, \ldots, X_p$. On note X le $(p+1)-$vecteur $(1, X_1, ..., X_p)$, $\widetilde{X}$ la matrice (de taille $n \times (p+1)$) des n observations de X et $\widetilde{Y} := (Y^{(1)}, \ldots, Y^{(n)})$ le vecteur des n observations de la variable aléatoire réponse Y. Le modèle suppose qu'on a la relation matricielle :*

$$\widetilde{Y} = \widetilde{X}\beta + \tilde{\epsilon}, \tag{1.2}$$

où $\beta \in \mathbb{R}^{p+1}$ est inconnu et supposé constant et $\tilde{\epsilon} := (\epsilon^{(1)}, \ldots, \epsilon^{(n)})$ est un vecteur aléatoire d'erreurs (non observé), c'est-à-dire que pour tout $i \in \{1, \ldots, n\}$, on a :

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)},$$

sous les hypothèses :

1. *les $\epsilon^{(i)}$ sont de moyenne nulle,*
2. *les $\epsilon^{(i)}$ ont même variance notée σ^2 (hypothèse d'homoscédasticité),*
3. *les $\epsilon^{(i)}$ sont indépendants,*
4. *les $\epsilon^{(i)}$ sont distribués selon une loi normale.*

En résumé, on a : $\tilde{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$.

Remarques :

— Dans le modèle linéaire classique, l'hypothèse 4. de normalité n'est pas requise et devient simplement "les $\epsilon^{(i)}$ sont distribués selon la même loi". En effet, l'estimation ponctuelle est possible sans aucune hypothèse de distribution sur les erreurs grâce à la méthode des moindres carrés présentée ci-dessous. En revanche, cette hypothèse de normalité des erreurs est primordiale pour l'estimation par intervalle de confiance et la construction de tests.

— Les hypothèses faites sur les erreurs impliquent que pour tout $i \in \{1, \ldots, n\}$, $Y^{(i)}$ est gaussienne, de moyenne $\beta_0 + \beta_1 X_1^{(i)} + \cdots + \beta_p X_p^{(i)}$ (d'où le nom de modèle linéaire) et de variance σ^2. De plus, ces $Y^{(i)}$ sont indépendants. Le modèle linéaire gaussien propose en fait une modélisation de la loi de $\widetilde{Y}$ par une loi gaussienne $\mathcal{N}_n(X\beta, \sigma^2 I_n)$.

— Parfois, on considère que les variables $X_1, \ldots, X_p$ sont aléatoires. Il faut alors rajouter l'hypothèse que la variable d'erreur ϵ est indépendante du vecteur $X = (1, X_1, \ldots, X_p)$.

Estimation du paramètre du modèle β
L'estimation du paramètre β (par la méthode des moindres carrés) est obtenue par :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} ||\widetilde{Y} - \widetilde{X}\beta||_2^2.$$

Il est appelé estimateur des moindres carrés et $\widetilde{X}\hat{\beta}$ correspond en fait exactement à la projection de $\widetilde{Y} \in \mathbb{R}^n$ sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $1, X_1, \ldots, X_p$ c'est-à-dire par les (vecteurs) colonnes de $\widetilde{X}$. Sous l'hypothèse que le rang de $\widetilde{X}$ est plein (donc égal à $p+1$), on obtient que $\hat{\beta} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y}$. Cet estimateur est le même que celui du maximum de vraisemblance lorsqu'on fait l'hypothèse de normalité des erreurs ϵ, et donc de Y.

Dans ce qui suit, on désignera souvent par abus de langage le modèle linéaire gaussien par régression linéaire ou modèle linéaire.

1.1.2 Généralités

Définition 1.1.2 (Lien direct)**.** *Soit X un p-vecteur aléatoire et soit $j, k \in \{1, \ldots, p\}$ tels que $j \neq k$. On dit qu'il y a un lien direct entre les variables X_j et X_k si les variables X_j et X_k sont dépendantes sachant les autres variables i.e. si $X_j \not\perp\!\!\!\perp X_k | (X_l)_{l \neq j,k}$.*

Le réseau qu'on cherche à inférer est un réseau de dépendances conditionnelles, *conditional dependency graph* en anglais [56, 45]. Il y a une arête entre les variables X_j et X_k s'il y a un lien direct entre elles :

$$X_j \longleftrightarrow X_k \iff X_j \not\perp\!\!\!\perp X_k | (X_l)_{l \neq j,k}.$$

Ce type de graphe est également appelé graphe de concentration ou graphe d'indépendance conditionnelle (totale).

Définition 1.1.3 (Corrélation conditionnelle)**.** *Soit X un p-vecteur aléatoire et soit $j, k \in \{1, \ldots, p\}$ tels que $j \neq k$. On définit la corrélation conditionnelle entre X_j et X_k sachant les autres variables $(X_l)_{l \neq j,k}$ comme :*

$$\mathrm{cor}(X_j, X_k \mid (X_l)_{l \neq j,k}) = \frac{\mathrm{cov}(X_j, X_k | \ (X_l)_{l \neq j,k})}{\sqrt{\mathrm{var}(X_j | \ (X_l)_{l \neq j,k}) \mathrm{var}(X_k | \ (X_l)_{l \neq j,k})}}.$$

Rappelons à présent un lemme d'algèbre linéaire, qui sera utile dans les résultats suivants :

Lemme 1.1.1 (Inverse d'une matrice par blocs [47])**.** *Si D est inversible ainsi que $A - BD^{-1}C$ (avec A, B, C et D des matrices de taille* ad hoc*), on a :*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \Big(A - BD^{-1}C\Big)^{-1} & -\Big(A - BD^{-1}C\Big)^{-1} BD^{-1} \\ -D^{-1}C\Big(A - BD^{-1}C\Big)^{-1} & D^{-1} + D^{-1}C\Big(A - BD^{-1}C\Big)^{-1} BD^{-1} \end{pmatrix}$$

Proposition 1.1.1. *Soit $X \sim \mathcal{N}_p(0, \Sigma)$ et soit (A, B) deux ensembles disjoints de $\{1, \ldots, p\}$. On suppose que Σ est définie positive.*
On a alors :

$$X_A | X_B \sim \mathcal{N}_{card(A)}\Big(\mu_{A|B}, \Sigma_{A|B}\Big),$$

où :

- $\mu_{A|B} = (\Sigma_{A,B}(\Sigma_{B,B})^{-1} X_B)'$
- $\Sigma_{A|B} = \Sigma_{A,A} - \Sigma_{A,B}(\Sigma_{B,B})^{-1}\Sigma_{B,A}$
- *Si M est une matrice de $\mathcal{M}_p(\mathbb{R})$, v un vecteur de $\mathbb{R}^p$ et I, J deux sous-ensembles de $\{1, \ldots, p\}$:*
  - *v_I est le sous-vecteur de v défini par $v_I := (v_i)_{i \in I}$*

- $M_{I,J}$ *est la sous-matrice de* M *définie par* $M_{I,J} := (m_{i,j})_{i\in I, j\in J}$

Démonstration. Il s'agit de calculer la densité conditionnelle de $X_A|X_B$ à l'aide de la densité d'un vecteur gaussien et de la formule d'inversion donnée dans le lemme 1.1.1. □

D'après la proposition 1.1.1, la loi du vecteur (X_j, X_k) sachant le reste est encore gaussienne et l'indépendance conditionnelle entre X_j et X_k sachant le reste est caractérisée par la nullité de la covariance conditionnelle et donc de la corrélation conditionnelle $\mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k})$:

$$\mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) = 0 \iff X_j \perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k}.$$

Ainsi, dans le modèle graphique gaussien :

$$X_j \longleftrightarrow X_k \iff \mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) \neq 0.$$

Cette correspondance n'est toutefois pas vraie en générale ! En fait, la corrélation conditionnelle de droite est en général aléatoire mais on va voir que ce n'est incroyablement pas le cas dans le modèle gaussien (corollaire 1.1.1).

1.1.3 Lien avec la matrice de précision

Définition 1.1.4 (Matrice de précision)**.** *Soit* $X \sim \mathcal{N}_p(0, \Sigma)$*. On appelle matrice de précision du vecteur* X *l'inverse de la matrice de covariance* Σ^{-1} *qu'on notera* K *par commodité.*

Proposition 1.1.2. *Soit* $X \sim \mathcal{N}_p(0, \Sigma)$ *et soit* (A, B) *deux ensembles disjoints de* $\{1, \dots, p\}$*. On suppose que* Σ *est définie positive.*
On a alors, avec les mêmes notations que la proposition 1.1.1 :

$$\left(\Sigma_{A|B}\right)^{-1} = \left[\Sigma^{-1}\right]_{A,A} = K_{A,A}.$$

Démonstration. Quitte à permuter les lignes et les colonnes de K et Σ, on peut supposer que $K = \begin{pmatrix} K_{A,A} & K_{A,B} \\ K_{B,A} & K_{B,B} \end{pmatrix}$ et $\Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix}$. Comme $K = \Sigma^{-1}$, en utilisant le lemme 1.1.1 d'inversion des matrices par blocs, on obtient :

$$\begin{aligned} K_{A,A} &= \left(\Sigma_{A,A} - \Sigma_{A,B}(\Sigma_{B,B})^{-1}\Sigma_{B,A}\right)^{-1} \\ &= \left(\Sigma_{A|B}\right)^{-1} \qquad \text{d'après la proposition 1.1.1.} \end{aligned}$$

□

Corollaire 1.1.1. *Soit* $j, k \in \{1, \dots, p\}$ *tels que* $j \neq k$ *:*

$$\mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) = -\frac{\left[\Sigma^{-1}\right]_{j,k}}{\sqrt{\left[\Sigma^{-1}\right]_{j,j}\left[\Sigma^{-1}\right]_{k,k}}} = -\frac{K_{j,k}}{\sqrt{K_{j,j}K_{k,k}}}.$$

En particulier,

$$X_j \perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k} \iff \left[\Sigma^{-1}\right]_{j,k} = 0 \ \left(= \left[\Sigma^{-1}\right]_{k,j}\right).$$

Remarques :

— En d'autres termes, de la même façon que la matrice de covariance d'un vecteur gaussien donne la structure d'indépendance d'un vecteur gaussien, la matrice de précision donne celle d'indépendance conditionnelle.

— La corrélation conditionnelle n'est donc pas aléatoire dans le modèle gaussien, ce qui est assez surprenant (et commode).

Démonstration. On applique la proposition précédente avec $A = \{j, k\}$ et $B = \{1, \dots, p\} \setminus A$. On a :

$$\begin{aligned} \operatorname{cor}(X_j, X_k \mid X_B) &= \frac{(\Sigma_{A|B})_{1,2}}{\sqrt{(\Sigma_{A|B})_{1,1}(\Sigma_{A|B})_{2,2}}}, && \text{par définition} \\ &= \frac{\left[(K_{A,A})^{-1}\right]_{1,2}}{\sqrt{\left[(K_{A,A})^{-1}\right]_{1,1}\left[(K_{A,A})^{-1}\right]_{2,2}}}, && \text{d'après la proposition 1.1.2.} \end{aligned}$$

Sans perte de généralité, supposons que $j < k$. On a :

$$(K_{A,A})^{-1} = \frac{1}{\det K_{A,A}} \begin{pmatrix} K_{j,j} & -K_{k,j} \\ -K_{j,k} & K_{k,k} \end{pmatrix}.$$

D'où : $\operatorname{cor}(X_j, X_k \mid (X_l)_{l \neq j,k}) = -\dfrac{K_{j,k}}{\sqrt{K_{j,j}K_{k,k}}}$. □

1.1.4 Lien avec les coefficients des régressions linéaires

Proposition 1.1.3. *Soit $X \sim \mathcal{N}_p(0, \Sigma)$. On suppose que Σ est définie positive. Soit $j \in \{1, \dots, p\}$. On effectue la régression linéaire de X_j sur les autres variables :*

$$X_j = \sum_{\substack{l=1 \\ l \neq j}}^{p} \beta_l^j X_l + \epsilon^j.$$

Alors on a : $\beta_l^j = -\dfrac{\left[\Sigma^{-1}\right]_{j,l}}{\left[\Sigma^{-1}\right]_{j,j}} = -\dfrac{K_{j,l}}{K_{j,j}}$ pour tout $l \neq j$.

Démonstration. Rappelons qu'on note K la matrice de précision. Tout d'abord, d'après le lemme 1.1.1, comme $K = \Sigma^{-1}$, on a :

$$\begin{aligned} K_{A,B} &= -\Big(\Sigma_{A,A} - \Sigma_{A,B}(\Sigma_{B,B})^{-1}\Sigma_{B,A}\Big)^{-1}\Sigma_{A,B}(\Sigma_{B,B})^{-1} \\ &= -\Sigma_{A|B}^{-1}\Sigma_{A,B}(\Sigma_{B,B})^{-1} && \text{d'après la proposition 1.1.1} \\ &= -K_{A,A}\Sigma_{A,B}(\Sigma_{B,B})^{-1} && \text{d'après la proposition 1.1.2.} \end{aligned}$$

D'où :

$$(K_{A,A})^{-1}K_{A,B} = -\Sigma_{A,B}(\Sigma_{B,B})^{-1} \tag{1.3}$$

D'après la proposition 1.1.1 avec $A = \{j\}$ et $B = \{1, \dots, p\} \setminus A$, on a que :

$$\mathbb{E}(X_j | X_l, l \neq j) = \Sigma_{j,B}(\Sigma_{B,B})^{-1}X_B.$$

Par ailleurs, on a que : $\mathbb{E}(X_j|X_l, l \neq j) = \sum_{\substack{l=1 \\ l\neq j}}^{p} \beta_l^j X_l$. Par identification, on obtient :

$$\begin{aligned} \sum_{\substack{l=1 \\ l\neq j}}^{p} \beta_l^j X_l &= \Sigma_{j,B}(\Sigma_{B,B})^{-1} X_B \\ &= -(K_{j,j})^{-1} K_{j,B} X_B \quad \text{d'après (1.3).} \end{aligned}$$

□

1.1.5 Lien avec les corrélations partielles

On a vu en introduction que les corrélations partielles présentent un certain intérêt dans l'inférence de réseaux. Cramér (1946) [25] en rappelle certaines propriétés. Elles sont ainsi définies :

Définition 1.1.5 (Corrélation partielle ou corrélations des résidus)**.** *Soit X un p-vecteur de moyenne nulle et admettant une matrice de covariance C définie positive et soit $j, k \in \{1, \dots, p\}$ tels que $j \neq k$. On effectue les régressions linéaires de X_j et X_k sur les autres variables X_l, $l \neq j, k$:*

$$\begin{aligned} X_j &= \sum_{\substack{l=1 \\ l\neq j,k}}^{p} \beta_l^j X_l + \epsilon^j, \\ X_k &= \sum_{\substack{l=1 \\ l\neq j,k}}^{p} \beta_l^k X_l + \epsilon^k, \end{aligned}$$

où les coefficients β sont tels que : $\mathbb{E}\Big(\big[X_j - \sum_{\substack{l=1 \\ l\neq j,k}}^{p} \beta_l^j X_l\big]^2\Big)$ et $\mathbb{E}\Big(\big[X_k - \sum_{\substack{l=1 \\ l\neq j,k}}^{p} \beta_l^k X_l\big]^2\Big)$ sont minimales. Alors la corrélation partielle entre X_j et X_k par rapport à $(X_l)_{l\neq j,k}$ est définie comme suit :

$$\rho(X_j, X_k \mid (X_l)_{l\neq j,k}) = \operatorname{cor}(\epsilon^j, \epsilon^k) = \mathbb{E}[\epsilon^j \epsilon^k].$$

À l'inverse des corrélations conditionnelles, ces corrélations partielles ne sont pas aléatoires. Elles correspondent à la corrélation entre X_j et X_k après avoir soustrait le meilleur estimateur linéaire à l'aide des variables restantes $(X_l)_{l\neq j,k}$, c'est-à-dire après avoir supprimé la part de variation due à l'influence de ces autres variables. Le résultat majeur les concernant est le suivant :

Proposition 1.1.4. *Soit X un p-vecteur de moyenne nulle et admettant une matrice de covariance C définie positive et soit $j, k \in \{1, \dots, p\}$ tels que $j \neq k$. Alors :*

$$\rho(X_j, X_k \mid (X_l)_{l\neq j,k}) = -\frac{\gamma_{j,k}}{\sqrt{\gamma_{j,j}\gamma_{k,k}}},$$

où $\gamma_{m,n}$ est le cofacteur d'indice m, n de la matrice de covariance $C = (c_{i,j})_{i,j\in\{1,\dots,p\}}$.

Corollaire 1.1.2. *Soit $X \sim \mathcal{N}_p(0, \Sigma)$ et soit $j, k \in \{1, \dots, p\}$ tels que $j \neq k$. On suppose que Σ est définie positive. Alors, les corrélations conditionnelles coïncident avec les corrélations partielles :*

$$\operatorname{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) = \rho(X_j, X_k \mid (X_l)_{l\neq j,k}).$$

Démonstration. Ici, $C = \Sigma$ et $K = \Sigma^{-1}$.

$$\begin{aligned}
\operatorname{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) &= -\frac{K_{j,k}}{\sqrt{K_{j,j}K_{k,k}}}, && \text{d'après le corollaire 1.1.1}\\
&= -\frac{\frac{\gamma_{j,k}}{\det C}}{\sqrt{\frac{\gamma_{j,j}}{\det C}\frac{\gamma_{k,k}}{\det C}}} \\
&= \rho(X_j, X_k \mid (X_l)_{l\neq j,k}), && \text{d'après la proposition 1.1.4.}
\end{aligned}$$

□

Avant de donner une preuve de la proposition 1.1.4, introduisons au préalable quelques notations :

- Soit $M = (m_{i,j})_{i,j\in\{1,\dots,p\}}$ une matrice de taille $p\times p$ à coefficients réels et soit A et B deux sous-ensembles de $[\![1,p]\!]$.
  - On note $M_{A,B} := (m_{i,j})_{i\in A,\ j\in B}$.
  - On note $M_{-A,-B} := (m_{i,j})_{i\in[\![1,p]\!]\setminus A,\ j\in[\![1,p]\!]\setminus B}$. Par abus de notation, si les ensembles sont des singletons, on s'affranchira des accolades en indice.
- On note $\Gamma = (\gamma_{i,j})_{i,j\in\{1,\dots,p\}}$ la matrice des cofacteurs de C : $\gamma_{i,j} = (-1)^{i+j}\det(C_{-i,-j})$. On rappelle que $C\Gamma' = \Gamma' C = \det C I_p$.
- Soit A et B deux sous-ensembles de $[\![1,p]\!]$ et $i\notin A$, $j\notin B$. On note $\gamma_{-A,-B;i,j}$ le cofacteur de la matrice $C_{-A,-B}$ correspondant aux variables X_i et X_j et $\Gamma_{-A,-B}$ la matrice correspondant à ces cofacteurs.

Démonstration. Sans perte de généralité, on le montre pour $\{j,k\} = \{1,2\}$.

Lemme 1.1.2. *Soit X un p-vecteur de moyenne nulle et admettant une matrice de covariance C définie positive. On effectue la régression linéaire de X_1 sur les autres variables :*

$$X_1 = \sum_{l=2}^{p}\beta_l^1 X_l + \epsilon^1_{2,3,\dots,p},$$

où les coefficients β sont tels que : $\mathbb{E}\Big(\big[X_1 - \sum_{l=2}^{p}\beta_l^1 X_l\big]^2\Big)$ est minimale.
Alors : $\beta_l^1 = -\dfrac{\gamma_{l,1}}{\gamma_{1,1}}$ pour tout $l\in[\![2,p]\!]$.

Démonstration. $\mathbb{E}\Big(\big[X_1 - \sum_{l=2}^{p}\beta_l^1 X_l\big]^2\Big)$ est minimale implique que le gradient par rapport aux coefficients β_l^1 est nul. Ainsi, $\forall i\in[\![2,p]\!]$:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_i^1}\mathbb{E}\Big(\big[X_1 - \sum_{l=2}^{p}\beta_l^1 X_l\big]^2\Big) = 0\\
\iff &\mathbb{E}\Big(\frac{\partial}{\partial\beta_i^1}[X_1 - \sum_{l=2}^{p}\beta_l^1 X_l]^2\Big) = 0, \text{ d'après le théorème de dérivation sous le signe somme}\\
\iff &\mathbb{E}\Big(-X_i[X_1 - \sum_{l=2}^{p}\beta_l^1 X_l]\Big) = 0\\
\iff &\sum_{l=2}^{p}\beta_l^1 c_{i,l} = c_{i,1}, \text{ car } X \text{ étant centré, } \mathbb{E}(X_iX_l) = c_{i,l}.
\end{aligned}$$

Donc : $C_{-1,-1}\begin{pmatrix}\beta_2^1\\ \vdots\\ \beta_n^1\end{pmatrix} = C_{-1,1}$. Comme C est définie positive, $C_{-1,-1}$ l'est aussi et ce système admet une unique solution :

$$\begin{pmatrix}\beta_2^1\\ \vdots\\ \beta_n^1\end{pmatrix} = (C_{-1,-1})^{-1}C_{-1,1} = \frac{1}{\gamma_{1,1}}\Gamma_{-1,-1}C_{-1,1}.$$

Ainsi, pour tout $i \in [\![21, p]\!]$:

$$\begin{aligned}
\beta_i^1 &= \frac{1}{\gamma_{1,1}}\sum_{k=2}^{p}\gamma_{-1,-1;i,k}c_{k,1}\\
&= \frac{1}{\gamma_{1,1}}\sum_{k=2}^{p}(-1)^{i-1+k-1}\det(C_{-\{1,i\},-\{1,k\}})c_{k,1}\\
&= \frac{1}{\gamma_{1,1}}\sum_{k=2}^{p}(-1)^{i-1+k-1}(-1)^{1+k-1}(-1)^{1+k-1}\det(C_{-\{1,i\},-\{1,k\}})c_{1,k}\\
&= (-1)^i\frac{1}{\gamma_{1,1}}\sum_{k=2}^{p}(-1)^{1+k-1}\det(C_{-\{1,i\},-\{1,k\}})c_{1,k}\\
&= (-1)^i\frac{\det C_{-i,-1}}{\gamma_{1,1}} = -\frac{\gamma_{i,1}}{\gamma_{1,1}}.
\end{aligned}$$

□

Lemme 1.1.3. *On a :*

- $\mathbb{E}[\epsilon^1_{2,3,\dots,p}] = 0$.
- $\mathrm{var}[\epsilon^1_{2,3,\dots,p}] = \mathbb{E}\Big([\epsilon^1_{2,3,\dots,p}]^2\Big) = \dfrac{\det C}{\gamma_{1,1}}$.

Démonstration. On a : $\epsilon^1_{2,3,\dots,p} = X_1 - \sum_{l=2}^{p}\beta_l^1 X_l$ et le vecteur X est centré. Par ailleurs,

$$\begin{aligned}
\mathbb{E}[X_i\epsilon^1_{2,3,\dots,p}] = \mathbb{E}[X_i\sum_{l=1}^{p}\frac{\gamma_{l,1}}{\gamma_{1,1}}X_l] = \sum_{l=1}^{p}\frac{\gamma_{l,1}}{\gamma_{1,1}}\mathbb{E}[X_iX_l] = \frac{1}{\gamma_{1,1}}\sum_{l=1}^{p}\gamma_{l,1}c_{i,l}\\
= \begin{cases}\dfrac{\det C}{\gamma_{1,1}} & \text{si } i = 1,\\ 0 & \text{sinon.}\end{cases}
\end{aligned}$$

Et $\mathrm{var}[\epsilon^1_{2,3,\dots,p}] = \mathbb{E}[\epsilon^1_{2,3,\dots,p} \times \epsilon^1_{2,3,\dots,p}] = \mathbb{E}[X_1\epsilon^1_{2,3,\dots,p}]$. □

On cherche ici à déterminer la corrélation partielle entre X_1 et X_2 sachant le reste :

$$\rho(X_1, X_2 \mid (X_l)_{l\geq 3}) = \mathrm{cor}(\epsilon^1_{3,4,\dots,p}, \epsilon^2_{3,4,\dots,p}) = \frac{\mathbb{E}[\epsilon^1_{3,4,\dots,p}\epsilon^2_{3,4,\dots,p}]}{\sqrt{\mathbb{E}\Big([\epsilon^1_{3,4,\dots,p}]^2\Big)\mathbb{E}\Big([\epsilon^2_{3,4,\dots,p}]^2\Big)}}.$$

D'une part, le vecteur $(X_1, X_3, \dots, X_p)$ est centré et a pour matrice de covariance $C_{-2,-2}$ (définie positive). Par analogie avec les lemmes 1.1.2 et 1.1.3, on a donc : $\epsilon^1_{3,4,\dots,p} = \sum_{l\neq 2}\dfrac{\gamma_{-2,-2;l,1}}{\gamma_{-2,-2;1,1}}X_l$

et $\mathbb{E}\Big([\epsilon^1_{3,4,\dots,p}]^2\Big) = \dfrac{\det C_{-2,-2}}{\gamma_{-2,-2;1,1}}$. De la même façon, on obtient : $\epsilon^2_{3,4,\dots,p} = \sum_{l\neq 1} \dfrac{\gamma_{-1,-1;l,2}}{\gamma_{-1,-1;2,2}} X_l$ et $\mathbb{E}\Big([\epsilon^2_{3,4,\dots,p}]^2\Big) = \dfrac{\det C_{-1,-1}}{\gamma_{-1,-1;2,2}}$. Remarquons que :

- $\det C_{-1,-1} = \gamma_{1,1}$ et $\det C_{-2,-2} = \gamma_{2,2}$.
- $\gamma_{-2,-2;1,1} = 1 \times \det(C_{[\![3,p]\!],[\![3,p]\!]}) = \gamma_{-1,-1;2,2}$.

Ainsi, $\mathbb{E}\Big([\epsilon^1_{3,4,\dots,p}]^2\Big) = \dfrac{\gamma_{2,2}}{\gamma_{-1,-1;2,2}}$ et $\mathbb{E}\Big([\epsilon^2_{3,4,\dots,p}]^2\Big) = \dfrac{\gamma_{1,1}}{\gamma_{-1,-1;2,2}}$.
Enfin :

$$\begin{aligned}
\mathbb{E}[\epsilon^1_{3,4,\dots,p}\epsilon^2_{3,4,\dots,p}] &= \mathbb{E}[\sum_{k\neq 2} \frac{\gamma_{-2,-2;k,1}}{\gamma_{-2,-2;1,1}} X_k \sum_{l\neq 1} \frac{\gamma_{-1,-1;l,2}}{\gamma_{-1,-1;2,2}} X_l] \\
&= \mathbb{E}[X_1 \epsilon^2_{3,4,\dots,p}] \\
&\quad \text{par le même argument que dans la preuve du lemme 1.1.3.} \\
&= \frac{1}{\gamma_{-1,-1;2,2}} \sum_{l\neq 1} \gamma_{-1,-1;l,2} c_{1,l} \\
&= \frac{\det C_{-1,-2}}{\gamma_{-1,-1;2,2}} \\
&\quad \text{en développant } \det C_{-1,-2} \text{ selon la première colonne.} \\
&= \frac{-\gamma_{1,2}}{\gamma_{-1,-1;2,2}}.
\end{aligned}$$

Finalement :

$$\begin{aligned}
\rho(X_1, X_2 \mid (X_l)_{l\geq 3}) &= \frac{\mathbb{E}[\epsilon^1_{3,4,\dots,p}\epsilon^2_{3,4,\dots,p}]}{\sqrt{\mathbb{E}\Big([\epsilon^1_{3,4,\dots,p}]^2\Big)\mathbb{E}\Big([\epsilon^2_{3,4,\dots,p}]^2\Big)}} \\
&= \frac{\dfrac{-\gamma_{1,2}}{\gamma_{-1,-1;2,2}}}{\sqrt{\dfrac{\gamma_{2,2}}{\gamma_{-1,-1;2,2}} \dfrac{\gamma_{1,1}}{\gamma_{-1,-1;2,2}}}} \\
&= -\frac{\gamma_{1,2}}{\sqrt{\gamma_{1,1}\gamma_{2,2}}}.
\end{aligned}$$

□

Le formule (voir Cramér (1946) [25]) donnée dans la proposition suivante permet d'exprimer la corrélation partielle conditionnée à un ensemble V en fonction de certaines corrélations partielles conditionnées à V privé d'une des variables. Ceci permet de diminuer la taille du conditionnement (appelé l'ordre) pour se ramener à des corrélations partielles d'ordres inférieurs.

Proposition 1.1.5. *Soit X un p-vecteur de moyenne nulle et admettant une matrice de covariance C définie positive. Soit $j, k \in \{1, \dots, p\}$ tels que $j \neq k$, $V \subset \{1, \dots, p\} \setminus \{j, k\}$ non vide et $i \in V$. Alors :*

$$\rho_{j,k|V} = -\frac{\rho_{j,k|V^*} - \rho_{i,j|V^*}\rho_{i,k|V^*}}{\sqrt{(1-\rho^2_{i,j|V^*})(1-\rho^2_{i,k|V^*})}},$$

où on note $\rho_{a,b|W} = \rho(X_a, X_b | X_w, w \in W)$ et $V^* = V \setminus \{i\}$.

Finalement, le cas gaussien est très sympathique. D'une part, les corrélations conditionnelles ne sont pas aléatoires et coïncident avec les corrélations partielles. Le cadre gaussien fournit également plusieurs caractérisations de l'indépendance conditionnelle entre X_j et X_k :

- La corrélation conditionnelle (et donc la corrélation partielle) est nulle :
$$\text{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) = \rho(X_j, X_k \mid (X_l)_{l\neq j,k}) = 0.$$
- Les coefficients correspondants dans la matrice de précision (qui est symétrique) sont nuls :
$$\left[\Sigma^{-1}\right]_{j,k} = \left[\Sigma^{-1}\right]_{k,j} = 0.$$
- Les coefficients de régression β_j^k (coefficient de X_j dans la régression de X_k sur les autres variables) et β_k^j (coefficient de X_k dans la régression de X_j sur les autres variables) sont nuls :
$$\beta_j^k = \beta_k^j = 0.$$

Ces différentes caractérisations fournissent différentes pistes pour l'inférence du réseau sous-jacent.

1.2 Approches statistiques pour l'inférence de réseaux

En pratique, différentes approches statistiques ont été proposées dans l'objectif d'inférer un réseau entre les variables $X_1, X_2, \dots X_p$. Rappelons que les données se présentent sous la forme :

$$X = \begin{pmatrix} X_1^{(1)} & \dots & X_p^{(1)} \\ X_1^{(2)} & \dots & X_p^{(2)} \\ & \dots & \\ X_1^{(n)} & \dots & X_p^{(n)} \end{pmatrix},$$

sous l'hypothèse que les observations $(X^{(i)})_{i=1,\dots,n}$ sont i.i.d. de loi $\mathcal{N}_p(\mu, \Sigma)$. Le graphe qu'on cherche à inférer est non-dirigé et contient une arête entre les variables X_j et X_k si et seulement si X_j et X_k sont dépendantes conditionnellement aux autres variables i.e. :

$$X_j \longleftrightarrow X_k \Longleftrightarrow X_j \not\perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k},$$

ce qui, dans le modèle gaussien, est équivalent à la non-nullité de la corrélation conditionnelle (et donc de la corrélation partielle associée) :

$$X_j \longleftrightarrow X_k \Longleftrightarrow \text{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) \neq 0.$$

Dans ce qui suit, nous allons présenter plusieurs approches pour estimer ce graphe. Certains outils de comparaison entre le graphe estimé et le graphe théorique sont récurrents. En voici un récapitulatif.

On dit qu'une arête estimée est :

- FP : *false positive*, si elle est détectée à tort, donc sans être dans le graphe théorique.
- FN : *false negative*, si elle n'est pas détectée et qu'elle n'est pas non plus dans le graphe théorique.
- TP : *true positive*, si elle est détectée correctement, donc en étant dans le graphe théorique.
- TN : *true negative*, si elle n'est pas détectée (à raison), c'est-à-dire sans être non plus dans le graphe théorique.

Voici à présent certains de ces outils de comparaisons :

- Le *False Discovery Rate* : $\mathrm{FDR} = \mathbb{E}\left[\frac{FP}{TP+FP}\right]$, qui représente la proportion moyenne de faux positifs parmi les estimés positifs.
- La sensibilité ou *True Positive Rate* TPR : $\mathrm{Sens} = \mathrm{TPR} = \frac{TP}{TP+FN}$, la proportion de vrais positifs parmi les positifs théoriques.
- La spécificité : $\mathrm{Spe} = \frac{TN}{TN+FP}$, la proportion de vrais négatifs parmi les négatifs théoriques.
- Le *False Positive Rate* FPR : $\mathrm{FPR} = \frac{FP}{TN+FP} = 1 - \mathrm{Spe}$, la proportion de faux positifs parmi les négatifs théoriques.

Villers *et al.* (2008) [99] donnent une synthèse complète et une comparaison des méthodes d'inférence publiées jusqu'alors et proposent de classer ces différentes approches en trois groupes selon la méthode qu'elles utilisent : estimation des matrices de covariance ou précision, approximation du graphe de concentration et estimation des voisinages.

1.2.1 Méthodes basées sur l'estimation des matrices de covariance ou précision

Les méthodes exposées ici visent à estimer la matrice de covariance ou la matrice de précision afin d'exploiter le résultat donné dans le corollaire 1.1.1 à savoir : la nullité des coefficients de la matrice de précision $K := \Sigma^{-1}$ fournit la structure d'indépendances conditionnelles entre les variables $X_1, \ldots, X_p$.

Une première approche naïve consiste à estimer la matrice de covariance par la matrice de covariance empirique $S = \frac{1}{n}(X - \bar{X})'(X - \bar{X})$ où $\bar{X}$ est la matrice $n \times p$ composée de n lignes identiques $(\bar{X_1}, \ldots, \bar{X_p})$ et de l'inverser. Il reste ensuite à seuiller S^{-1} pour décider de la nullité des coefficients. Le problème est que la matrice de covariance empirique est souvent mal conditionnée voire non inversible (notamment quand le nombre d'observations n est plus petit que le nombre de variables p), conduisant à une très mauvaise estimation de la matrice de précision quand cela est possible. Dempster (1972) [28] et Cox et Wermuth (1996) [24] exposent les problèmes liés à la *covariance selection*. En 2005, Schäfer et Strimmer [86] combinent deux idées : une estimation bootstrap plus stable de la matrice de covariance (*bagging*) permettant de réduire sa variance et la méthode de pseudo-inversion de Moore-Penrose.

La même année, Schäfer et Strimmer [87] proposent d'utiliser un estimateur shrinkage pour la matrice de covariance qui consiste en une combinaison linéaire de la matrice de covariance empirique S et d'un estimateur cible $\hat{\Omega}$ de variance faible : $\hat{\Sigma} = \lambda\hat{\Omega} + (1-\lambda)S$. Le paramètre λ se calcule directement à partir des données et le plus souvent $\hat{\Omega} = I$ ou $\text{diag}(S)$. On a donc $\hat{K} = (\hat{\Sigma})^{-1}$.

Pour chacune des méthodes présentées dans ces deux articles, il reste ensuite à décider quelles entrées sont considérées comme non nulles. Schäfer et Strimmer supposent que les entrées de la matrice de précision estimée suivent un modèle mixte connu dont les paramètres sont à estimer. Ceci leur permet de construire un test basé sur l'estimation des probabilités d'existence de chaque arête pour déterminer quelles entrées sont significativement différentes de zéro. On peut remarquer que cette méthode, contrairement aux méthodes exposées par la suite, même si elle est ici utilisée pour le modèle graphique gaussien, pourrait également servir à estimer les corrélations partielles, proportionnelles aux coefficients de la matrice de précision, dans un modèle plus général.

Par la suite, des méthodes ont été présentées dans la littérature, utilisant le Lasso introduit en 1996 par Tibshirani [94] dans le cadre de la régression linéaire et permettant de regrouper à la fois les étapes d'estimation de la matrice de covariance/précision et la spécification des entrées non-nulles. Cette méthode offre une solution attractive aux problèmes de mauvais conditionnements (particulièrement liés au faible nombre d'observations) et consiste en l'ajout d'une pénalisation L_1 dans l'estimation des coefficients de la matrice de précision $K = \Sigma^{-1}$. Ceci permet d'annuler les coefficients relatifs aux interactions les moins importantes, rendant la matrice solution parcimonieuse (*sparse* en anglais) et faisant du Lasso un des critères de pénalisation les plus populaires (d'abord dans un contexte de régression [119, 77, 92, 45]). Ces méthodes sont davantage détaillées dans la sous-section 2.1.2 du chapitre 2.

Huang *et al.* (2006) [48] présentent une méthode non-paramétrique pour estimer la matrice de covariance dans le modèle graphique gaussien en réécrivant la matrice de covariance grâce à une décomposition de Cholesky modifiée.

Friedman *et al.* (2008) [36] et Banerjee *et al.* (2008) [6] proposent d'estimer directement la matrice de précision grâce à la vraisemblance L_1-pénalisée du modèle gaussien. En effet, la log-vraisemblance du modèle gaussien est : $\log(\det K) - \text{trace}(SK)$. Ainsi, il s'agit de résoudre le problème d'optimisation suivant sur l'ensemble des matrices Θ définies positives ($\Theta \succ 0$) :

$$\underset{\Theta \succ 0}{\text{argmax}} \ \log(\det \Theta) - \text{trace}(S\Theta) - \lambda||\Theta||_1,$$

où $||.||_1$ est la norme L_1 c'est-à-dire la somme des valeurs absolues des coefficients et $\lambda > 0$ le paramètre de pénalisation. D'autres auteurs comme Dahl *et al.* (2008) [26] ou Yuan et Lin (2007) [117] se sont également intéressés à ce problème d'optimisation qu'ils ont résolu, comme [6], de façon exacte en adaptant des méthodes de points intérieurs. [6] et [36] résolvent ce problème grâce à un algorithme de descente par blocs (block coordinate descent algorithm) ; la différence est que [6] utilise des méthodes de points intérieurs (IPM) lors d'une des étapes de l'algorithme alors que la procédure graphical Lasso ou Glasso de [36] utilise une résolution Lasso sur un problème dual. Finalement, la procédure Glasso consiste à résoudre récursivement des problèmes de régression Lasso, ce qui rend l'algorithme plus performant. Pour un paramètre de pénalisation $\lambda > 0$ donné, cette résolution permet de fournir une estimation $\hat{K}(\lambda)$ de la matrice de précision (l'algorithme de [6] retourne la matrice de covariance) et les arêtes du graphe estimé sont donc les (i, j) tels

que $\hat{K}(\lambda)_{i,j} \neq 0$.

Banerjee *et al.* [6] proposent de plus un choix de paramètre $\lambda(\alpha)$ pour lequel la probabilité de lier deux composantes connexes du graphe théorique est bornée par α. La procédure Glasso est implémentée dans le package R éponyme et dans le package R `huge`.

1.2.2 Méthodes basées sur l'estimation d'une approximation du graphe de concentration

Les méthodes mentionnées ici n'estiment pas le graphe d'indépendance conditionnelle totale (graphe de concentration) mais un graphe qui s'en approche. On rappelle que ces méthodes sont utilisées dans le modèle gaussien pour lequel les notions de corrélations partielles (déterministes) et corrélations conditionnelles (aléatoires) coïncident.

Notamment, Wille et Bühlmann (2006) [111] approchent le graphe d'indépendance conditionnelle totale par le graphe d'indépendance conditionnelle 0-1. Au lieu de regarder les corrélations conditionnelles sachant toutes les autres variables restantes, on regarde les corrélations conditionnelles jusqu'au premier ordre, c'est-à-dire les corrélations simples et les corrélations conditionnelles sachant chacune des autres variables une à une, c'est-à-dire $\mathrm{cor}(X_j, X_k|X_l)$ pour $l \in \{1, \ldots, p\} \setminus \{j, k\}$. L'avantage de cette procédure est qu'elle permet d'estimer ces quantités convenablement même sur un faible nombre d'observations tout en reflétant raisonnablement le graphe de concentration. Les corrélations conditionnelles (partielles) d'ordre 1 sont déterminées à partir des corrélations simples en utilisant la formule liant les corrélations partielles (proposition 1.1.5) à savoir :

$$\rho(X_j, X_k|X_l) = -\frac{\mathrm{cor}(X_j, X_k) - \mathrm{cor}(X_i, X_j)\mathrm{cor}(X_i, X_k)}{\sqrt{(1 - \mathrm{cor}^2(X_i, X_j))(1 - \mathrm{cor}^2(X_i, X_k))}},$$

pour $j, k, l \in \{1, \ldots, p\}$ distincts. Ils infèrent alors un réseau qu'ils appellent *0-1 conditional independence graph* où il y a une arête entre X_j et X_k si la corrélation classique et toutes les corrélations $\mathrm{cor}(X_j, X_k|X_l)_{l \neq j,k}$ sont non nulles. En pratique, pour déterminer s'il y a une arête entre X_j et X_k, ils utilisent le test du rapport de vraisemblance pour tester la nullité de chacune de ces corrélations. Les p-valeurs obtenues sont ensuite corrigées par des procédures de tests multiples comme Bonferroni ou Benjamini-Hochberg. Pour un risque fixé α, ils montrent que la probabilité de détecter une arête à tort (FP) est plus petite que α pour un nombre d'observations n assez grand. Cette méthode présente des propriétés computationnelles intéressantes mais les graphes résultants coïncident rarement avec les graphes de concentration théoriques, ce qui est notamment dû à un fort nombre de faux négatifs.

Castela et Roverato (2006) [16] ainsi que Malouche et Sevestre (2007) [65] étendent cette procédure à l'ordre q et construisent le graphe q-partiel pour $0 \leq q \leq p - 2$. Ce graphe contient une arête entre X_j et X_k si toutes les corrélations partielles conditionnées à un ensemble de cardinal q sont non nulles. Ils établissent des conditions sous lesquelles ce graphe partiel coïncide avec le graphe de concentration total (c'est-à-dire le graphe partiel d'ordre $p-2$). Castela et Roverato considèrent ainsi toutes les corrélations partielles $\rho(X_j, X_k|X_L)$ où $L \subset \{1, \ldots, p\} \setminus \{j, k\}$, $\mathrm{card}(L) = q$ et $X_L = \{X_l\}_{l \in L}$ et testent leur nullité à l'aide d'un taux de non-rejet. Malouche et Sevestre utilisent une procédure itérative pour estimer le graphe partiel à partir du graphe partiel estimé du précédent ordre.

Kalisch et Buhlmann (2007) [52], eux, proposent d'inférer le squelette d'un graphe orienté acyclique. Ce graphe, appelé aussi graphe d'indépendance conditionnelle forte, est une approximation beaucoup plus restrictive consistant à mettre une arête entre X_j et X_k si et seulement si toutes les corrélations conditionnelles (donc partielles) entre X_j et X_k sachant n'importe quel groupe de variables restantes sont non nulles. On voit aisément que ce graphe est un sous-ensemble du graphe d'indépendance conditionnelle totale. L'estimation de ce type de graphe est réalisée à l'aide d'une procédure itérative rapide initialement utilisée pour les graphes orientés acycliques parcimonieux, l'algorithme PC (pour *partial correlation*). L'idée est de partir du graphe plein et de retirer une arête dès que la corrélation partielle correspondante est nulle. On commence par calculer les corrélations partielles d'ordre 0 (les corrélations simples), puis on calcule pas à pas les corrélations d'ordre supérieur correspondant à des arêtes encore présentes grâce à la formule récursive donnée à la proposition 1.1.5. La nullité de chaque corrélation est testée à chaque étape grâce à un test basé sur la transformation de Fisher.

1.2.3 Méthodes basées sur l'estimation des voisinages

Une première méthode est la méthode Lasso de Meinshausen et Bühlmann (2006) [71]. Celle-ci repose sur le résultat de la proposition 1.1.3 explicitant le lien entre les coefficients des régressions linéaires et la matrice de précision dans le modèle gaussien. Ainsi, ils proposent d'estimer le voisinage $ne(k) := \{X_j, j \in \{1, \dots, p\} \setminus \{k\} : \Sigma^{-1}_{k,j} \neq 0\}$ de chaque sommet X_k (c'est-à-dire de chaque variable) à l'aide d'une estimation Lasso des coefficients de la régression linéaire. L'estimation des coefficients de la régression linéaire s'obtient traditionnellement par la méthode des moindres carrés. Plus précisément, les auteurs décident ici d'ajouter une pénalisation L_1 dans l'estimation de ces coefficients pour forcer un nombre restreint de coordonnées à être non nulles et donc rendre le voisinage plus parcimonieux. Ainsi, pour un sommet X_k fixé, on estime les coefficients de la régression linéaire de X_k sur les autres variables $X_l, l \neq k$ par :

$$\hat{\beta}^k(\lambda) = \underset{\beta^k \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \Big(||X_k - \sum_{j \neq k} \beta^k_j X_j||^2_2 + \lambda ||\beta^k||_1 \Big),$$

pour un paramètre de pénalisation $\lambda > 0$ à déterminer. Le voisinage estimé $\hat{ne}_\lambda(k)$ du sommet X_k est alors :

$$\hat{ne}_\lambda(k) := \{X_j, j \in \{1, \dots, p\} \setminus \{k\} : \hat{\beta}^k_j(\lambda) \neq 0\}.$$

À partir de ces estimations, deux graphes peuvent être construits : le graphe "or" qui consiste à mettre une arête entre les sommets X_j et X_k si l'un des deux coefficients $\hat{\beta}^j_k$ et $\hat{\beta}^k_j$ est non nul et le graphe "and" où l'arête existe si ces deux coefficients sont non nuls.

Dans cet article, Meinshausen et Bühlmann établissent deux principaux résultats. Le premier concerne les erreurs de type I, c'est-à-dire d'inclure à tort des voisins dans l'estimation du voisinage (les faux positifs). Sous certaines conditions et pour un choix de λ qui décroît vers 0 à un taux plus petit que $n^{-1/2}$ (où n est la taille de l'échantillon), $\mathbb{P}(\hat{ne}_\lambda(k) \subset ne(k))$ croît vers 1 exponentiellement vite avec n. Le second concerne les erreurs de type II, c'est-à-dire d'oublier des voisins dans l'estimation du voisinage (les faux négatifs). Sous les mêmes conditions, ils montrent que $\mathbb{P}(ne(k) \subset \hat{ne}_\lambda(k))$ croît vers 1 exponentiellement vite. Au vu de ces deux résultats, on conclut que, sous certaines hypothèses, la méthode est consistante et la probabilité que les voisinages estimés et théoriques coïncident converge vers 1 avec la taille n de l'échantillon.

En pratique, il est plus intéressant de privilégier la version "and" si on veut limiter le nombre de faux positifs, c'est-à-dire les arêtes détectées à tort. Cette version est plus spécifique et permet de rendre le FDR (*false discovery rate*, i.e. la proportion de faux positifs parmi les positifs

estimés), plus petit que la version "or", moins sélective. Cette procédure est implémentée dans le package R `huge`.

Giraud et ses coauteurs (2009) [40] proposent une méthode mêlant les avantages de différentes procédures : la régularisation (pénalisation) L_1, le critère de sélection de modèle BIC et un critère de sélection précédemment établi par Giraud (2008) [39]. Il s'agit d'une procédure en deux temps qui construit d'abord une famille de graphes candidats puis qui applique une procédure de sélection sur ces graphes. Ces graphes sont supposés parcimonieux dans la mesure où on se restreint à des graphes de degrés inférieurs à $n-2$.

Concernant l'élaboration d'une famille de graphes candidats, les auteurs fixent 4 familles de graphes candidats basés sur des résultats théoriques annexes et des études de simulation. Il est conseillé d'utiliser l'une d'entre elles ou éventuellement l'union de deux d'entre elles pour éviter que le critère de sélection de la seconde partie soit trop compliqué à mettre en œuvre. Les familles candidates sont :

— La famille C01 basée sur la procédure d'estimation de Wille et Bühlmann (2006) [111] sur le graphe d'indépendance conditionnelle 0-1.

— La famille Lasso-And basée sur la procédure d'estimation de Meinshausen et Bühlmann (2006) [71] et sur l'algorithme Lasso LARS [32].

— La famille Lasso adaptif, une version modifiée de la famille précédente inspirée du Lasso adaptif de Zou [121].

— La famille quasi-exhaustive basée sur la minimisation d'un critère sur tous les graphes de degré maximal fixé.

Cette étape conduit à une famille $\mathcal{G}$ de graphes potentiellement convenables, caractérisés par leurs arêtes. Il reste maintenant à choisir un de ces graphes.

On note θ la matrice associée au graphe qui correspond aux coefficients des régressions linéaires, c'est-à-dire θ est la matrice $p \times p$ de diagonale nulle telle que :

$$\mathbb{E}_\Sigma[X_j | X_l, l \neq j] = \sum_{l \neq j} \theta_{j,l} X_l.$$

En fait, on a $\theta_{j,l} = -\dfrac{\Sigma^{-1}_{j,l}}{\Sigma^{-1}_{j,j}}$ (d'après la proposition 1.1.3). Pour chaque graphe G de la famille $\mathcal{G}$, on associe un estimateur $\hat{\theta}_G$ de θ :

$$\hat{\theta}_G = \underset{\theta' \in \Theta_G}{\operatorname{argmin}} \, ||X(I - \theta')||,$$

où $||.||$ est la norme de Frobenius et Θ_G est l'ensemble des matrices convenables, c'est-à-dire les matrices $p \times p$ telles que le coefficient (j,l) est non-nul si et seulement s'il y a une arête entre X_j et X_l dans le graphe G. Ceci revient à estimer les coefficients par la méthode classique des moindres carrés mais en ayant au préalable imposé les arêtes. Pour chaque estimateur $(\hat{\theta}_G)_{G \in \mathcal{G}}$, on calcule un certain critère dépendant du nombre de voisins de chaque variable et d'une fonction de pénalisation. Le "meilleur" graphe est le graphe associé à l'estimateur $\hat{\theta}_G$ qui minimise ce critère.

Des garanties théoriques sont données dans un cadre non-asymptotique, notamment concernant le Mean Square Error of Prediction (MSEP), qui est une quantité utilisée pour mesurer la qualité de l'estimateur $\hat{\theta}$ final. Sous l'hypothèse que les degrés des graphes G de la famille $\mathcal{G}$ soient inférieurs à une quantité de l'ordre de $\frac{n}{2} \log p$, MSEP$(\hat{\theta})$ est majoré par une quantité

proche de $\min_{G \in \mathcal{G}} MSEP(\hat{\theta}_G)$ à un facteur $\log p$ près. Les auteurs de cet article donnent également un résultat asymptotique de consistance de la procédure sous certaines conditions.

En pratique, il s'avère que cette méthode donne de bonnes performances surtout quand n est petit, mais est difficilement réalisable quand $p > 40$ à cause de la difficulté computationnelle notamment liée au choix des graphes.

Chapitre 2

Modèles binaires

Sommaire

2.1 Régression logistique 21
2.1.1 Généralités 21
2.1.2 Modèle linéaire généralisé et pénalisations 22
2.2 Le modèle d'Ising 25
2.2.1 Brève présentation du modèle d'Ising 25
2.2.2 Lien avec la régression logistique 26
2.2.3 Simulation de données suivant la loi d'Ising 27
2.2.4 Extensions du modèle d'Ising 29
2.3 Approches statistiques 31
2.3.1 Sur la régression logistique pénalisée 31
2.3.2 Dans le cas de l'inférence de réseaux 31
2.3.3 Comparaisons pratiques des méthodes Glasso et régression logistique 32

Dans ce chapitre, on s'intéresse toujours à l'inférence de réseaux mais dans un cadre binaire. En vue d'étudier des données inflatées en zéro, une manière d'inférer le réseau pourrait être de transformer les données de manière binaire en absence/présence ou nul/non-nul. Ces données peuvent donc être modélisées par des variables de Bernoulli. On présente d'abord quelques résultats théoriques sur la régression logistique et le modèle d'Ising avant d'en expliquer les applications à l'inférence de réseaux. On en profite pour présenter le modèle linéaire généralisé, dont la régression logistique et la régression linéaire sont des cas particuliers, ainsi que les principaux types de pénalisations.

2.1 Régression logistique

2.1.1 Généralités

Le modèle de régression logistique est un modèle où la variable à expliquer est binaire (voir [110, 2] ou le cours de Laurent Rouvière [85]). C'est par exemple le cas lorsqu'on s'intéresse à l'absence/présence d'une maladie, aux intentions de vote ou encore aux résultats d'un sondage à question fermée.

Définition 2.1.1 (Régression logistique)**.** *Soit Y une variable aléatoire à valeurs dans $\{0,1\}$ à expliquer par p variables explicatives $X_1,\dots,X_p$ et notons $X = (1, X_1, \dots, X_p)'$. Le modèle logistique propose une modélisation de la loi de Y par une loi de Bernoulli de paramètre*

$p_\beta(x) = \mathbb{P}_\beta(Y = 1|X = x)$ *telle que :*

$$\text{logit } p_\beta(x) = \log \frac{p_\beta(x)}{1 - p_\beta(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x'\beta, \tag{2.1}$$

logit désignant la fonction bijective et dérivable de]0,1[dans $\mathbb{R}$ *:* $p \mapsto \log(\frac{p}{1-p})$*. L'égalité* (2.1) *peut également s'écrire :*

$$p_\beta(x) = \mathbb{P}_\beta(Y = 1|X = x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

Remarques :

— Dans un modèle logistique, nous effectuons deux choix pour définir le modèle :

1. Le "choix" d'une loi de Bernoulli pour $Y|X$ (dans le modèle linéaire, on fait le choix d'une loi normale). Il ne s'agit en réalité par d'un vrai choix dans ce cas dans la mesure où Y est binaire.
2. Le choix de la modélisation de $\mathbb{P}_\beta(Y = 1|X = x)$ par $\text{logit}(\mathbb{P}_\beta(Y = 1|X = x)) = x'\beta$. La fonction logit est appelée fonction de lien.

Remarquons également que : $\mathbb{E}_\beta[Y|X = x] = p_\beta(x)$ et $\mathbb{V}_\beta[Y|X = x] = p_\beta(x)(1 - p_\beta(x))$, ce qui implique que la variance n'est pas constante et varie en fonction de x (contrairement au modèle gaussien où on fait souvent l'hypothèse d'homoscédasticité).

— La fonction de lien aurait pu être différente (il existe plusieurs fonctions définies sur $]0, 1[$ ayant une forme sigmoïdale) mais logit a l'avantage d'être symétrique et plus simple à manipuler. Par ailleurs, elle présente des avantages en terme d'interprétation d'odds (de chances) : pour tout $x \in \mathbb{R}^p$, $\text{odd}(x) = \dfrac{\mathbb{P}_\beta(Y = 1|X = x)}{\mathbb{P}_\beta(Y = 0|X = x)} = \exp(x'\beta)$ et d'odds ratio : pour tout $x, \tilde{x} \in \mathbb{R}^p$, $\dfrac{\text{odd}(x)}{\text{odd}(\tilde{x})} = \exp((x - \tilde{x})'\beta)$

— On peut interpréter ce modèle à l'aide d'une variable latente. Si on introduit ϵ une variable aléatoire symétrique et une variable latente Y^* non observée : $Y^* = \tilde{\beta}_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ telle que $Y|X = x$ vaut 1 si Y^* est plus grande qu'un seuil s et 0 sinon, on obtient alors :

$$\mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y^* > s) = \mathbb{P}(-\epsilon < \tilde{\beta}_0 - s + \beta_1 x_1 + \cdots + \beta_p x_p) = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p),$$

où F est la fonction de répartition de ϵ et $\beta_0 = \tilde{\beta}_0 - s$. La régression logistique correspond à la loi logistique pour ϵ, mais d'autres lois sont envisageables (par exemple, une loi normale conduit au modèle probit).

— Il s'agit d'un cas particulier du modèle linéaire généralisé (voir ci-dessous). Ainsi, plusieurs aspects sont similaires, comme l'estimation des paramètres par la méthode du maximum de vraisemblance. La log-vraisemblance à maximiser est $\sum_{i=1}^{n} [y^{(i)} \log(p_\beta(x^{(i)})) + (1 - y^{(i)}) \log(1 - p_\beta(x^{(i)}))]$ où $y^{(i)}$ et $x^{(i)}$ correspondent respectivement à la i-ème observation de Y et de X.

2.1.2 Modèle linéaire généralisé et pénalisations

La régression linéaire et la régression logistique sont toutes deux des cas particuliers des modèles linéaires généralisés [104, 67, 2, 107].

Définition 2.1.2 (Famille exponentielle)**.** *On dit qu'une loi appartient à la famille exponentielle si elle admet une densité f_θ par rapport à une mesure μ qui peut s'écrire sous la forme :*

$$f_\theta(x) = a(x)b(\theta)\exp\left(T(x).Q(\theta)\right),$$

où θ est le paramètre de la loi (éventuellement multidimensionnel).

La plupart des lois usuelles appartiennent à la famille exponentielle, par exemple, lois de Poisson, lois Gamma, lois normales, lois binomiales, loi de Bernoulli (par rapport à la mesure de comptage : $a(x) = 1$, $b(p) = 1 - p$, $T = id$ et $Q(p) = \log(\frac{p}{1-p})$)...

Définition 2.1.3 (Modèle linéaire généralisé)**.** *On est dans un contexte de régression, c'est-à-dire qu'on dispose d'une variable aléatoire réponse Y et de p variables explicatives $X = (X_1, \ldots, X_p)$ (aussi appelées covariables). Le but est d'exprimer l'espérance de la variable réponse Y en fonction d'une combinaison linéaire $X\beta$ des variables explicatives. Un modèle linéaire généralisé comporte trois composantes :*

- *la composante aléatoire : on suppose que les lois des n observations $Y^{(i)}$ du modèle sont dominées par une même mesure et appartiennent à la famille exponentielle,*
- *le prédicteur linéaire qui correspond à la quantité $X\beta$,*
- *la fonction de lien g : $g(\mathbb{E}[Y|X]) = X\beta$.*

Les régressions linéaire et logistique sont bien des cas particuliers du modèle linéaire généralisé pour les fonctions de lien $g = id$ et $g = \text{logit}$ respectivement. L'estimation des paramètres $\beta = (\beta_j)_{j=1,\ldots,p}$ du modèle se fait par maximisation de la log-vraisemblance de l'échantillon $(Y^{(i)}, X^{(i)})_{i=1,\ldots,n}$.

Ce type de régression a d'abord été essentiellement utilisé pour faire de la prédiction, puis dans le but de faire de la sélection de variables pour faciliter l'interprétation, c'est-à-dire de déterminer quels prédicteurs sont réellement liés à la variable réponse et jouent un rôle important. La sélection de variables consiste à déterminer la nullité des coefficients $(\beta_k)_{k=1,\ldots,p}$ associés aux p variables explicatives. Naturellement, le Lasso et d'autres types de pénalisations ont été introduits dans ce but mais également dans le but de faire face aux problèmes liés à la grande dimension, comme les problèmes de conditionnement. Les chercheurs se sont d'abord penchés sur l'introduction d'une pénalisation dans la régression linéaire qui consiste à minimiser :

$$C_{\lambda,\gamma}(\beta) := ||Y - X\beta||^2 + \lambda||\beta||_\gamma^\gamma,$$

où $\lambda \geq 0$, $0 \leq \gamma \leq 2$ et $||\beta||_\gamma^\gamma = \begin{cases} \sum_{k=1}^{m} \mathbb{1}_{\beta_k \neq 0}, \text{ si } \gamma = 0, \\ \sum_{k=1}^{m} |\beta_k|^\gamma, \text{ sinon.} \end{cases}$.

Par dualité lagrangienne, à $\lambda \geq 0$ fixé, ceci est équivalent à minimiser $||Y - X\beta||^2$ sous la contrainte $||\beta||_\gamma^\gamma \leq \tau$ pour un certain $\tau \geq 0$, $\lambda = 0$ correspondant à un τ infini et vice versa. $C_{\lambda,\gamma}$ n'est convexe que si $\gamma \geq 1$ et présente des propriétés de parcimonie (*sparsity* en anglais) pour λ assez grand (et donc τ assez petit) que si $\gamma \leq 1$. Cette pénalisation se généralise ensuite à l'estimation des coefficients dans les modèles linéaires généralisés en remplaçant $||Y - X\beta||^2$

par $L(\beta)$, la log-vraisemblance du modèle.

La pénalisation correspondant à $\gamma = 0$ porte le nom de "sélection de modèle". La sélection de modèle consiste à minimiser un critère pénalisé par le degré du modèle sur l'ensemble des 2^p modèles possibles. Quand le nombre p de covariables est grand, cette approche devient cependant vite insoluble.
La pénalisation L_2 ($\gamma = 2$), appelée Ridge et introduite par Hoerl et Kennard (1988) [46], a d'abord été utilisée pour pallier ces problèmes de dimension. La pénalisation Ridge offre en effet plusieurs avantages : elle permet notamment de résoudre les problèmes de conditionnement et d'estimer un modèle quand les variables explicatives sont fortement corrélées et consiste à résoudre un problème d'optimisation strictement convexe qui fournit une solution explicite dans le modèle linéaire gaussien. Toutefois, elle ne conduit pas à des solutions parcimonieuses et n'est pas adaptée pour la sélection automatique de variables.
Le Lasso, ou pénalisation L_1 ($\gamma = 1$), a été introduite plus récemment par Tibshirani (1996) [94]. Cette pénalisation conduit à une solution parcimonieuse (si le paramètre de pénalisation λ est assez grand) tout en gardant la convexité du problème d'optimisation. C'est en fait la plus petite valeur γ pour laquelle le problème d'optimisation reste convexe, ce qui simplifie grandement la résolution du problème d'optimisation, de même que la parcimonie de la solution. Ceci explique en grande partie la notoriété du Lasso, particulièrement adapté pour la sélection de variables. Les propriétés théoriques du Lasso ont depuis été considérablement analysées, par exemple par Donoho et Elad (2003) [29] et Tropp (2006) [96] qui ont notamment étudié les raisons pour lesquelles cette pénalisation mène à des solutions parcimonieuses, et plus récemment par Zhao et Yu (2006) [119] ou Zou *et al.* (2007) [123].

Beaucoup de variantes dérivées du Lasso se sont développées, la liste suivante étant loin d'être exhaustive (voir Hastie *et al.* (2015) [45] pour plus de détails).

- Le *relaxed Lasso* (Meinshausen (2007) [70]) est une procédure en deux temps qui utilise le Lasso pour déterminer un sous-ensemble $E \subset \{1, \dots, p\}$ de variables potentiellement impliquées puis une ré-estimation ordinaire des coefficients de régression β dont le support est inclus dans E.
- *elastic net* (Zou et Hastie (2005) [122]) est un compromis entre la pénalisation Ridge et la pénalisation Lasso, qui consiste à remplacer $||\beta||_\gamma^\gamma$ par une combinaison convexe entre la norme L_1 et la norme L_2 : $(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1$, $0 \leq \alpha \leq 1$. Le Lasso, contrairement à Ridge, ne se comporte notamment pas très bien par rapport aux variables fortement corrélées. Ceci permet de combiner les avantages de la pénalisation Lasso et de la pénalisation Ridge : obtenir une solution parcimonieuse tout en gérant les problèmes de structure de corrélations des covariables.
- Le *group Lasso* (Yuan et Lin (2006) [116], Meier *et al.* (2008) [69]) fait l'hypothèse que les covariables possèdent une structure naturelle de groupes. Il permet ainsi de tenir compte de cette structure en supposant que le coefficient de régression est le même pour les variables d'un même groupe et de sélectionner ou éliminer simultanément des variables d'un même groupe.
- L'*overlap group Lasso* (Jacob *et al.* (2009) [49]) est une extension du *group Lasso* qui autorise une variable à appartenir à plusieurs groupes.
- Le *hierarchical group Lasso* (Zhao *et al.* (2009) [118], Lim et Hastie (2014) [60]) est une utilisation du overlap group Lasso qui permet d'introduire une hiérarchie sur les variables (si on veut en sélectionner une avant une autre par exemple).

— Le *fused Lasso* (Tibshirani *et al.* (2005) [95]) permet de prendre en compte la spatialité des variables. Il comporte deux termes de pénalisation : le premier étant une pénalisation Lasso pour rendre la solution parcimonieuse, et le second pour encourager les coefficients des variables voisines à être proches.

— L'*adaptive Lasso* (Zou (2006) [121]) a été introduit pour pallier le biais considérable que peuvent avoir les estimateurs des grands coefficients. Il consiste à ajouter des poids dépendant des données dans la pénalisation : $\sum_{k=1}^{p} \mathbf{w_k}|\beta_k|$.

Des packages R ont été développés pour les modèles linéaires généralisés, incluant également différents types de pénalisation : le package `glmnet` [37] qui propose différentes pénalisations comme le Lasso, Ridge ou *elastic net* ou encore le package `glmpath` [77].

Le Lasso constitue à ce jour une des méthodes les plus utilisées pour la sélection de variables. Il reste ensuite le choix délicat du paramètre de pénalisation λ. Même si des techniques classiques peuvent être utilisées pour calibrer ce paramètre de pénalisation, ce réglage demeure un problème ouvert important.

2.2 Le modèle d'Ising

2.2.1 Brève présentation du modèle d'Ising

Le modèle introduit ci-dessous est souvent appelé *modèle d'Ising* dans la littérature de mécanique statistique ; dans la littérature du machine learning, il est plus connu sous le nom de *Boltzmann machines*.

Définition 2.2.1. *Soit Θ une matrice symétrique de $\mathcal{M}_p(\mathbb{R})$.*
On dit qu'un p-vecteur aléatoire X à valeurs dans $\{0,1\}^p$ suit une loi d'Ising [44, 6, 38, 82] de paramètre Θ si :

$$\forall x \in \{0,1\}^p,\ \mathbb{P}_\Theta(X = x) = \pi_\Theta(x) = \frac{\exp\Big(\sum_{i=1}^{p} \Theta_{ii} x_i + \sum_{i<j} \Theta_{ij} x_i x_j\Big)}{Z(\Theta)}, \tag{2.2}$$

où $Z(\Theta) = \sum_{x\in\{0,1\}^p} \exp\Big(\sum_{i=1}^{p} \Theta_{ii} x_i + \sum_{i<j} \Theta_{ij} x_i x_j\Big)$ est la constante de normalisation. On note $X \sim \mathcal{I}_p(\Theta)$.

Remarques : Dans la littérature, il arrive qu'on trouve des alternatives, notamment :

— Chaque composante du vecteur peut prendre deux valeurs, parfois il s'agit plutôt de $\{-1;1\}$.

— Il arrive qu'on ait :

$$\begin{aligned}\mathbb{P}_\Theta(X = x) = \pi_\Theta(x) &= \frac{\exp(\sum_{i=1}^{p} \Theta_{ii} x_i + \sum_{i\neq j} \Theta_{ij} x_i x_j)}{Z(\Theta)} \\ &= \frac{\exp(\sum_{i=1}^{p} \Theta_{ii} x_i + 2\sum_{i<j} \Theta_{ij} x_i x_j)}{Z(\Theta)},\end{aligned}$$

ce qui revient à diviser par 2 les coefficients hors diagonale de la matrice symétrique paramètre Θ pour obtenir les mêmes probabilités qu'avec (2.2).

— Si $p = 1$, $\mathcal{I}_p(\Theta)$ correspond à la loi de Bernoulli de paramètre $\dfrac{\exp(\Theta)}{1+\exp(\Theta)}$.

Théorème 2.2.1. *Soit $X \sim \mathcal{I}_p(\Theta)$ et soit $\{A, B\}$ une partition de $\{1, ..., p\}$. Alors $X_A|X_B$ suit encore une loi d'Ising de paramètre la matrice $\widetilde{\Theta}$ telle que : $\widetilde{\Theta}_{aa} = \Theta_{aa} + \sum_{b\in B} \Theta_{ab}x_b$ pour tout $a \in A$ et $\widetilde{\Theta}_{aa'} = \Theta_{aa'}$ pour tout $a, a' \in A$ distincts.*

Démonstration. Il suffit de calculer la densité de $X_A|X_B$, $\pi_{A|B}(x_A) = \dfrac{\pi_\Theta(x_A, x_B)}{\sum_{x_A\in\{0,1\}^{|A|}} \pi_\Theta(x_A, x_B)}$ □

Théorème 2.2.2. *Soit $X \sim \mathcal{I}_p(\Theta)$. L'indépendance conditionnelle de deux variables est donnée par la nullité du coefficient correspondant de la matrice Θ :*

$$X_j \perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k} \iff \Theta_{jk} = 0.$$

Démonstration. • On le fait d'abord pour $p = 2$.
Dans ce cas, comme X_1 et X_2 suivent des lois de Bernoulli, elles sont indépendantes si et seulement si :

$$\begin{aligned}
&\mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0)\\
\iff\quad &\frac{1}{Z(\Theta)} = \frac{(1+\exp(\Theta_{22}))(1+\exp(\Theta_{11}))}{Z(\Theta)^2}\\
\iff\quad &1 = \frac{1+\exp(\Theta_{11})+\exp(\Theta_{22})+\exp(\Theta_{11}+\Theta_{22})}{1+\exp(\Theta_{11})+\exp(\Theta_{22})+\exp(\Theta_{11}+\Theta_{22}+\Theta_{12})}\\
\iff\quad &\Theta_{12} = 0.
\end{aligned}$$

• D'après le théorème 2.2.1, la loi conditionnelle de (X_j, X_k) sachant le reste est une loi d'Ising de paramètre $\widetilde{\Theta} = \begin{pmatrix} \Theta_{jj} + \sum_{l\neq j,k} \Theta_{jl}x_l & \Theta_{jk} \\ \Theta_{jk} & \Theta_{kk} + \sum_{l\neq j,k} \Theta_{kl}x_l \end{pmatrix}$. D'après le cas $p = 2$, on conclut que X_j et X_k sont indépendantes conditionnellement aux $(X_l)_{l\neq j,k}$ si et seulement si $\widetilde{\Theta}_{12} = 0$ c'est-à-dire $\Theta_{jk} = 0$.

□

2.2.2 Lien avec la régression logistique

Théorème 2.2.3. *Soit $X \sim \mathcal{I}_p(\Theta)$ et soit $j \in \{1, \dots, p\}$. Notons $X_{-j} = (X_k)_{k\neq j}$. Alors la loi de $X_j|X_{-j}$ est une Bernoulli de paramètre $\text{logit}^{-1}(\Theta_{jj} + \sum_{k\neq j} \Theta_{kj}X_k)$ c'est-à-dire de paramètre $p_\beta(X_{-j})$ où $\beta = (\Theta_{jj}, \Theta_{1j}, \dots, \Theta_{j-1,j}, \Theta_{j+1,j}, \dots, \Theta_{pj})$ avec les notations de la section 2.1.*

Démonstration. Ceci découle du théorème 2.2.1. En effet, le paramètre de la Bernoulli est donné par $\pi_j(1|x_{-j})$ où $\pi_j(x_j|x_{-j}) = \dfrac{\exp\left(x_j(\Theta_{jj} + \sum_{k\neq j}\Theta_{jk}x_k)\right)}{\sum_{x_j\in\{0,1\}}\exp\left(x_j(\Theta_{jj} + \sum_{k\neq j}\Theta_{jk}x_k)\right)} = \dfrac{\exp\left(x_j(\Theta_{jj} + \sum_{k\neq j}\Theta_{jk}x_k)\right)}{1+\exp\left(\Theta_{jj} + \sum_{k\neq j}\Theta_{jk}x_k\right)}$. □

En d'autres termes, si $X \sim \mathcal{I}_p(\Theta)$, déterminer les indépendances conditionnelles deux à deux revient à effectuer la régression logistique de chacune des variables sur les autres et d'étudier la nullité des coefficients de ces régressions. Plus précisément, si on note $(\beta_k^j)_{\substack{k=0,\dots,p,\\ k\neq j}}$ les coefficients de la régression logistique de X_j sur X_{-j}, alors $\beta_k^j = \beta_j^k = \Theta_{kj} = \Theta_{jk}$ pour tout $k \in \{1,\dots,p\}\setminus\{j\}$. Ceci fournit un modèle pour l'inférence de réseaux sur des données binaires : on suppose qu'elles sont issues d'un modèle d'Ising et on effectue les régressions logistiques de chaque variables sur les autres, de façon similaire à ce qui est fait dans le modèle gaussien avec les régressions linéaires.

2.2.3 Simulation de données suivant la loi d'Ising

On se donne une matrice $\Theta \in \mathcal{M}_p(\mathbb{R})$ symétrique. La question est de savoir comment simuler X suivant la loi d'Ising de paramètre Θ. Calculer toutes les probabilités relatives aux 2^p configurations possibles $x \in \{0,1\}^p$ du p-vecteur X puis effectuer une méthode de rejet est beaucoup trop compliqué. Il existe en revanche plusieurs méthodes basées sur la méthode de Monte-Carlo par chaînes de Markov (MCMC) [76, 38]. Ces méthodes reposent sur la construction d'une chaîne de Markov $(Y_n)_{n\in\mathbb{N}}$ π_Θ-irréductible, apériodique et de matrice de transition P telle que la loi π_Θ de X est P-invariante (de sorte que Y_n converge en loi vers π_Θ).

Échantillonnage de Gibbs

La méthode d'échantillonnage de Gibbs (*Gibbs sampling*) consiste à partir d'un vecteur déterministe puis à simuler tour à tour chaque coordonnée à l'aide de la loi conditionnelle. Cette méthode repose sur la théorie des chaînes de Markov (et impose certaines conditions sur la chaîne qui ne sont pas toujours faciles à vérifier).

Algorithme de Gibbs

- Étape 1 : On commence avec une valeur initiale $x^0 \in \{0,1\}^p$.
- Étape 2 : À chaque itération l, on met à jour le vecteur x^l à l'aide des lois conditionnelles sachant l'itéré courant x^l. On choisit une coordonnée $k \in \{1,\dots,p\}$ puis on simule X_k selon $\pi_k(.|x^l_{-k})$ la loi conditionnelle de $X_k|X_{-k} = x^l_{-k}$.
- Étape 3 : On réitère l'étape 2 jusqu'à convergence.
- Étape 4 : Retourner l'itéré courant.

Dans l'étape 2, on peut faire plusieurs choix pour la coordonnée à mettre à jour. Notamment, deux cas classiques sont le balayage linéaire et le balayage aléatoire des coordonnées du vecteur. Le problème majeur est le critère de convergence de l'étape 3 (souvent un choix plus ou

moins arbitraire du nombre nécessaire d'itérations). On peut par exemple représenter les histogrammes des lois conditionnelles et évaluer graphiquement quand ceux-ci semblent rester stables.

Dans notre modèle, les lois conditionnelles sont :

— Dans le cas $\{0,1\}$, la loi de x_k sachant le reste est $\pi_k(0|x_{-k}) = \dfrac{1}{1+\exp(\theta_{kk}+\sum_{i\neq k}\theta_{ki}x_i)}$ et $\pi_k(1|x_{-k}) = 1 - \pi_k(0|x_{-k}) = \dfrac{1}{1+\exp(-(\theta_{kk}+\sum_{i\neq k}\theta_{ki}x_i))}$.

— Dans le cas $\{-1,1\}$, la loi de x_k sachant le reste est $\pi_k(1|x_{-k}) = \dfrac{1}{1+\exp(-2(\theta_{kk}+\sum_{i\neq k}\theta_{ki}x_i))}$ et $\pi_k(-1|x_{-k}) = 1 - \pi_k(1|x_{-k}) = \dfrac{1}{1+\exp(2(\theta_{kk}+\sum_{i\neq k}\theta_{ki}x_i))}$.

Metropolis-Hastings

La méthode de Metropolis-Hastings est également une méthode MCMC. L'idée est de choisir une valeur y sur $\{0,1\}^p$ et de remplacer l'itération courante x^l par y si cette valeur est plus probable. Plus précisément, on considère le rapport des deux probabilités $\alpha = \dfrac{\pi(y)}{\pi(x^l)}$: si $\alpha > 1$, on remplace l'itération courante par y, sinon, "on laisse une chance à y" et on remplace l'itération courante x^l par y avec probabilité α.

Algorithme de Metropolis-Hastings

– Étape 1 : On commence avec une valeur initiale $x^0 \in \{0,1\}^p$.
– Étape 2 : À chaque itération l :
  - On tire un y uniformément sur $\{0,1\}^p$.
  - On accepte y (c'est-à-dire $x^{l+1} = y$) avec probabilité $\alpha = \min(1, \dfrac{\pi(y)}{\pi(x^l)})$. Sinon $x^{l+1} = x^l$.
– Étape 3 : On réitère l'étape 2 jusqu'à convergence.
– Étape 4 : Retourner l'itéré courant.

Il s'agit ici d'un cas particulier. On pourrait choisir différemment y, notamment de façon dépendante de l'itération courante x^l à l'aide d'une matrice de transition stochastique (qui intervient alors dans la probabilité avec laquelle on choisit de garder y ensuite). L'échantillonneur de Gibbs est en fait un cas particulier de l'algorithme de Metropolis-Hastings pour la matrice de transition donnée par les probabilités conditionnelles $\pi_k(.|x^l_{-k})$.

Comme pour l'échantillonnage de Gibbs, le problème majeur est le critère de convergence de l'étape 3, qui consiste également souvent en un nombre d'itérations nécessaires à fixer pour assurer la convergence vers la loi désirée.

Exact sampling ou simulation parfaite

Le principal avantage de cette méthode est la convergence vers la loi exacte. En effet, contrairement aux deux algorithmes précédents dont le nombre nécessaire d'itérations est assez flou, la condition pour la convergence ici est une condition d'arrêt, ce qui assure la convergence vers la loi exacte [80].

L'idée est de partir de deux états "extrémaux" (on choisira $(1,\dots,1)$ et $(0,\dots,0)$) et d'itérer jusqu'à ce que les deux états coïncident (la condition d'arrêt). Pour les itérations, on utilise la méthode d'échantillonnage de Gibbs et la mise à jour doit se faire avec les mêmes variables uniformes. L'algorithme est le suivant :

Algorithme de simulation parfaite

– Étape 1 : On fixe M un nombre d'itérations et on commence avec $x^M = (0,\dots,0)$ et $y^M = (1,\dots,1)$.
– Étape 2 : Pour $l = M-1,\dots,0$:
  - On simule U^l un vecteur de p variables uniformes.
  - On met à jour les itérés courants x^{l+1} et y^{l+1} à l'aide des lois conditionnelles des itérés actualisés. Pour chaque coordonnée $k \in \{1,\dots,p\}$, on simule x_k^l et y_k^l selon $\pi_k(.|(x_1^l,\dots,x_{k-1}^l,x_{k+1}^{l+1},\dots,x_p^{l+1}))$ et $\pi_k(.|(y_1^l,\dots,y_{k-1}^l,y_{k+1}^{l+1},\dots,y_p^{l+1}))$ respectivement, en utilisant la même variable uniforme U_k^l. On obtient alors les nouveaux itérés x^l et y^l.
– Étape 3 : Tant que $x^0 \neq y^0$ (x et y n'ont pas coalescé) :
  - On simule M p-vecteur de variables uniformes : $U_{2M-1},\dots,U_{M+1},U_M$.
  - On double le nombre d'itérations M.
  - On réitère l'étape 2 en partant de x^M et y^M.
– Étape 4 : Retourner $x^0 (= y^0)$.

Remarque : Quand on double le nombre M d'itérations à effectuer, les M p-vecteurs de lois uniformes utilisés précédemment pour les M dernières itérations sont gardés.

Le désavantage majeur de cette méthode est en contrepartie son coût computationnel.

2.2.4 Extensions du modèle d'Ising

Le modèle d'Ising ne tient compte que des interactions deux à deux. Plusieurs auteurs tels que Dai, Ding et Wahba (2013) [27] et Loh et Wainwright (2013) [64] ont proposé une extension théorique permettant de prendre en compte les interactions d'ordre supérieur.

Le modèle de Bernoulli multivarié proposé par [27] est le suivant :

Soit Y un p-vecteur à valeurs dans $\{0,1\}^p$ et soit $y \in \{0,1\}^p$. La loi du vecteur Y est donnée par :

$$\mathbb{P}(Y=y) = \frac{\exp\Big(\sum_{r=1}^{p} \sum_{1\leq j_1<j_2<\dots<j_r\leq p} f^{j_1,\dots,j_r} y_{j_1}\dots y_{j_r}\Big)}{K(\mathbf{f})},$$

où $K(\mathbf{f})$ est la constante de renormalisation et $\mathbf{f}$ la famille de paramètres du modèle.

Plusieurs résultats sont similaires à ceux du modèle d'Ising :

— C'est une loi de la famille exponentielle (voir définition 2.1.2).

— Les lois marginales et conditionnelles sont encore des Bernoullis multivariées.

— Dans le cas où $p = 2$, X_1 et X_2 sont indépendantes si et seulement $f^{1,2} = 0$ (le cas $p = 2$ coïncide en fait avec le modèle d'Ising car il n'y a pas d'interaction d'ordre supérieur à 2).

— On a également une caractérisation de l'indépendance conditionnelle par rapport aux autres variables :

$$X_j \perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k} \Longleftrightarrow f^A = 0,\ \forall A \subset \{1,\dots,p\},\ \text{tel que } j \in A \text{ et } k \in A.$$

En revanche, ce modèle comporte $2^p - 1$ paramètres contre $\dfrac{p(p-1)}{2}$ pour le modèle d'Ising, ce qui le rend très difficile à mettre en pratique. D'ailleurs, les auteurs de cet article n'y exposent que des résultats théoriques. Notamment, l'inférence pour ce modèle n'est pas étudiée ni proposée et reste un peu ambiguë. En particulier, si on ne s'intéresse qu'à l'indépendance conditionnelle deux à deux, le modèle semble assez lourd computationnellement.

Remarque : Si on ne regarde que les interactions d'ordre au plus deux, on a une bijection entre la matrice d'adjacence, donc le graphe d'indépendance conditionnelle, et la nullité/non nullité des coefficients du modèle. C'est le cas des modèles gaussiens (avec la matrice de précision) et Ising (avec la matrice paramètre Θ). Ce sont en fait des cas où la nullité des coefficients du modèle caractérise l'indépendance conditionnelle.
En revanche, ce n'est plus le cas pour les modèles dans lesquels on s'intéresse à des interactions d'ordre supérieur. Typiquement, dans le modèle Bernoulli multivarié, il y a une arête entre X_j et X_k (donc un 1 en position (i,j) et (j,i) dans la matrice d'adjacence) si et seulement si il existe A tel que $j \in A$ et $k \in A$ tel que $f^A \neq 0$, ce qui n'est clairement pas une bijection (même en terme de nullité/non nullité des coefficients). Le même graphe pourrait donc être issu de différents modèles Bernoullis multivariés où les coefficients nuls/non nuls sont différents ; ceci fournit toutefois des informations supplémentaires pour l'interprétation des résultats.

Loh et Wainwright [64], quant à eux, proposent un modèle plus général basé sur une décomposition de la vraisemblance en fonction des cliques du graphe (les sous-ensembles de sommets complètement connectés). Pour un vecteur X de variables binaires dans $\{0,1\}$, la vraisemblance s'écrit :

$$f_\theta(x_1,\dots,x_p) = \exp\Big\{ \sum_{C\in\mathcal{C}} \theta_C \mathbb{1}_C(x_C) - \phi(\theta)\Big\},$$

où $\mathcal{C}$ est l'ensemble des cliques du graphe, $\mathbb{1}_C(x_C) := \prod_{s\in\mathcal{C}} x_s$ la statistique associée à la clique C, $\theta_C \in \mathbb{R}$ est le poids associé à la clique C et $\phi(\theta)$ la constante de renormalisation. Dans le cas d'interactions deux à deux, ce modèle revient au modèle classique d'Ising. Une extension à des variables discrètes à valeurs dans $\{0,\dots,m-1\}$ est proposée, consistant à associer à une clique C la famille de statistiques suivantes :

$$\mathbb{1}_{C,J}(x_C) = \begin{cases} 1, \text{ si } x_C = J, \\ 0, \text{ sinon.} \end{cases}$$

où J est une configuration possible pour les éléments de la clique telle qu'aucun des éléments ne soit nul (c'est-à-dire un élément de $\mathcal{X}_p^{|C|} := \{1, ..., m-1\}^{|C|}$) et $\theta_{C,J}$ correspond à la famille de poids associés. La vraisemblance devient alors :

$$f_\theta(x_1, \ldots, x_p) = \exp\{\sum_{C \in \mathcal{C}} \sum_{J \in \mathcal{X}_m^{|C|}} \theta_{C,J} \mathbb{1}_{C,J}(x_C) - \phi(\theta)\}.$$

2.3 Approches statistiques

2.3.1 Sur la régression logistique pénalisée

De nombreux auteurs se sont penchés sur la régression logistique et plus particulièrement sur la régression logistique pénalisée, motivée par le problème de sélection de variables. Ce sont en effet des questions, en marge de la prédiction et de l'interprétation, qui se posent naturellement dans différents domaines. Notamment, depuis plus d'un demi-siècle, la régression logistique Lasso est très populaire dans la recherche biomédicale.

D'un point de vue applicatif, Zhu et Hastie (2004) [120] l'ont par exemple utilisée dans un contexte de diagnostique du cancer afin de comprendre quels gènes sont impliqués ; Whu *et al.* (2009) [113] dans un contexte d'analyse du génome. Ceci permet entre autres de restreindre les études des biologistes à des sous-ensembles de gènes plus petits, ces études étant souvent chronophages et onéreuses.
Lang (1995) [55] l'a appliquée à la classification automatique de ressources documentaires par genre, par thème, par opinion etc.

En parallèle, de nombreux auteurs en ont étudié les propriétés théoriques et proposé différents algorithmes pour résoudre le problème d'optimisation Lasso. Parmi ces algorithmes, certains sont plutôt classiques dans la résolution de problèmes d'optimisation convexes comme IRLS (Lee *et al.* (2006) [59]) ou les méthodes de points intérieurs (Koh *et al.* (2007) [54]). Le package `glmnet` [37] utilise un algorithme de descente par coordonnée couplé à une approche de Newton itérative proximale alors que l'algorithme du package `glmpath` [77] est plutôt basé sur une méthode prédicteurs/erreurs d'optimisation convexe.

2.3.2 Dans le cas de l'inférence de réseaux

On est cette fois dans un contexte d'inférence de réseaux. Comme exposé précédemment, on dispose de données qu'on présente sous la forme d'une matrice X comme suit :

$$X = \begin{pmatrix} X_1^{(1)} & \ldots & X_p^{(1)} \\ X_1^{(2)} & \ldots & X_p^{(2)} \\ & \ldots & \\ X_1^{(n)} & \ldots & X_p^{(n)} \end{pmatrix},$$

où les observations, c'est-à-dire chaque ligne $i \in \{1, \ldots, n\}$ de la matrice X, sont supposées indépendantes. Dans ce cas, chaque $X^{(i)}$ prend ses valeurs dans $\{0, 1\}^p$. Le but est alors d'inférer un graphe d'indépendance conditionnelle entre les variables $X_1, X_2, \ldots, X_p$. Dans les approches

statistiques exposées par la suite, on fait l'hypothèse que $X^{(i)} \sim \mathcal{I}(\Theta)$ pour tout $i \in \{1, \ldots, n\}$.

Wainwright *et al.* (2007) [100] et Ravikumar *et al.* (2010) [82] utilisent une approche similaire à celle de Meinshausen et Bühlmann dans le modèle graphique gaussien [71]. Ils effectuent les régressions logistiques L_1-pénalisées de chaque variable sur les variables restantes et utilisent la parcimonie des coefficients de régression pour inférer le réseau sous-jacent. Les régressions logistiques pénalisées successives permettent en fait de déterminer la nullité des coefficients de la matrice paramètre dans un modèle d'Ising afin de déterminer les dépendances conditionnelles pour établir le graphe.

Sous certaines conditions théoriques, impliquant notamment la matrice d'information de Fisher des graphes à degrés bornés, et pour un paramètre de régularisation λ_n supérieur à une quantité de l'ordre de $\sqrt{\frac{\log p}{n}}$, ils prouvent que leur méthode est consistante. En fait, ils prouvent qu'avec probabilité de l'ordre de $1 - \exp(-\lambda_n^2 n)$, les voisinages estimés excluent correctement les faux positifs et incluent correctement les arêtes pour lesquelles le coefficient Θ_{ij} associé est assez grand. En d'autres termes, leur procédure exclut les faux positifs et les faux négatifs correspondent à des coefficients Θ_{ij} petits.

Ils illustrent ensuite leurs résultats avec des simulations en utilisant l'algorithme proposé par [54] basé sur des méthodes de points intérieurs.

La même année, Lee *et al.* [58] utilisent une procédure de gradient conjugué pour une maximisation exacte de la log-vraisemblance pénalisée dans le cas de modèles log-linéaires. Ils appliquent cependant cette méthode en pratique sur des données Ising.

En 2011, Jalali et ses coauteurs [51] poursuivent le travail entamé par [100] et [82] en utilisant un algorithme gourmand (*greedy* en anglais) qui montre des performances intéressantes à la fois sur le plan théorique et sur le plan expérimental.

En 2008, Banerjee et ses coauteurs [6] montrent que le problème d'estimation du graphe entre des variables binaires Ising peut être transformé en un problème équivalent à celui du graphical Lasso dans le modèle graphique gaussien présenté dans le chapitre précédent. Ils appliquent ainsi l'algorithme développé dans ce même article au problème d'inférence sur données binaires et montrent que pour un niveau α fixé, et le paramètre de pénalisation $\lambda_n(\alpha) = \dfrac{\chi^2(1 - \alpha/2p^2)}{\sqrt{n} \min\limits_{i>j} \hat{\sigma}_i \hat{\sigma}_j}$ (où $\hat{\sigma}_i$ est l'écart-type empirique de X_i et $\chi^2(\alpha)$ le quantile d'ordre α de la loi du chi-deux à 1 degré de liberté), la probabilité qu'il existe un faux positif est inférieure à α.

Ils appliquent ensuite leur méthode à l'inférence d'un graphe sur un ensemble de votes de p sénateurs américains afin de détecter des dépendances entre les politiciens. Il en résulte que la plupart des démocrates ont des voisins démocrates et la plupart des républicains ont des voisins républicains.

2.3.3 Comparaisons pratiques des méthodes Glasso et régression logistique

Dans ce qui suit, nous allons présenter quelques comparaisons entre les méthodes de Banerjee [6] (Glasso) et de Ravikumar [82]. On représentera la version "and" et la version "or" pour les résultats obtenus par la méthode de régression logistique (voir [82]).

Pour ce faire, les données sont simulées selon un modèle d'Ising de sorte que le réseau soit une grille de taille $d \times d$. On a choisi $d = 10$, il y a donc $p = 100$ variables. La matrice paramètre Θ comporte des coefficients non nuls lorsque les variables relatives sont liées ; ces coefficients sont tirés selon une loi de Rademacher $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$.

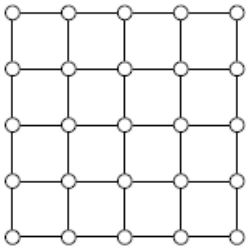


FIGURE 2.1 – Exemple de réseau "grille" pour $d = 5$.

Les données Ising sont ensuite simulées selon la méthode d'échantillonnage de Gibbs avec $n = 100$ et $n = 500$ observations. La méthode Glasso est effectuée à l'aide du package R `glasso` et les régressions logistiques pénalisées avec le package R `glmnet`. On représente les résultats sous la forme d'une courbe ROC qui consiste à représenter la sensibilité en fonction de 1 - spécificité pour différentes valeurs du paramètre de pénalisation λ_n.

Banerjee [6] propose un choix optimal de λ_n dépendant des données, pour un niveau α fixé, $\lambda_n(\alpha) = \dfrac{\chi^2(1-\alpha/2p^2)}{\sqrt{n}\min\limits_{i>j}\hat{\sigma}_i\hat{\sigma}_j}$ où $\hat{\sigma}_i$ est l'écart-type empirique de X_i et $\chi^2(\alpha)$ le quantile d'ordre α de la loi du chi-deux à 1 degré de liberté. On note par la suite $\lambda_n^B = \lambda_n(0.05)$, qui est le paramètre de pénalisation utilisé dans les simulations de l'article de Banerjee [6]. Ravikumar [82] préconise lui, un choix de λ_n indépendant des données mais dépendant de la taille n de l'échantillon et du nombre p de variables, $\lambda_n^R = \sqrt{\dfrac{\log p}{n}}$. L'ensemble des λ choisis pour établir les courbes ROC varie entre $\lambda_n^B/2$ et $\frac{3}{2}\lambda_n^B$ et $\lambda_n^R/2$ et $2\lambda_n^R$ respectivement ; on indique avec des symboles spécifiques les points correspondant aux valeurs de λ_n^R et λ_n^B.

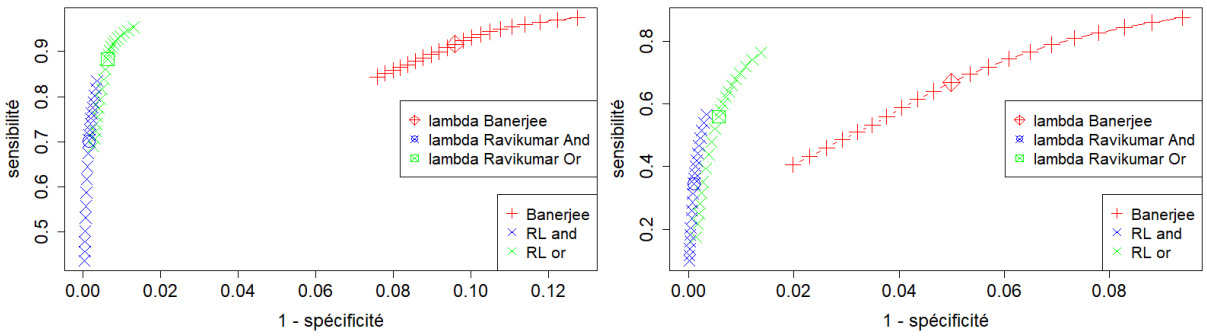
Commentaires : Plusieurs choses sont à commenter :

— Sans grande surprise, la taille de l'échantillon compte et les résultats sont meilleurs pour $n = 500 = 5p$, parfois considérablement en comparaison de $n = 100 = p$.

— Le choix du paramètre de pénalisation n'est globalement pas mauvais.Les méthodes semblent finalement ne détecter aucune arête d'où les résultats.

— D'une manière générale, on peut constater que la régression logistique "and" est plus spécifique et moins sensible que la régression logistique "or". La régression logistique est bien plus spécifique que la procédure Glasso mais également moins sensible que Glasso. La régression logistique est plus sensible aux variations du paramètre de pénalisation concernant la sensibilité alors que la procédure Glasso l'est davantage concernant la spécificité (pour une amplitude plus faible toutefois).

Pour conclure, la régression logistique semble conseillée dans un contexte où on privilégie la spécificité, c'est-à-dire quand on cherche à limiter le nombre de faux positifs (les arêtes détectées à tort). Il faut alors avoir en tête que beaucoup d'arêtes ne sont pas détectées, mais les arêtes



(a) $n = 500$. (b) $n = 100$.

FIGURE 2.2 – Courbe ROC pour données Ising simulées selon la méthode d'échantillonnage de Gibbs pour $n = 500$ et $n = 100$ échantillons. Les courbes ROC sont obtenues en moyennant les sensibilités et spécificités obtenues sur 100 répétitions.

détectées sont essentiellement des vraies arêtes. Si en revanche, le but est d'obtenir un ensemble d'arêtes contenant les vraies arêtes, il est plus judicieux d'utiliser la procédure Glasso en gardant à l'esprit qu'un certain nombre d'arêtes détectées sont des faux positifs.

Principales contributions

L'inférence de réseaux a de plus en plus d'applications notamment en santé humaine et en environnement pour l'étude de données micro-biologiques et génomiques. Les réseaux constituent en effet un outil approprié pour représenter, voire étudier des relations entre des entités. On a vu dans les chapitres précédents que de nombreuses techniques mathématiques d'estimation ont été développées notamment dans le cadre des modèles graphiques gaussiens (chapitre 1) mais aussi dans le cas de données binaires ou mixtes (chapitre 2).

Le traitement des données d'abondance (par exemple, de micro-organismes comme les bactéries) est particulier pour deux raisons : d'une part ces données ne reflètent pas toujours directement la réalité car un processus de séquençage a lieu pour dupliquer les espèces et ce processus apporte de la variabilité, d'autre part une espèce peut être absente dans certains échantillons. On est alors dans le cadre de données inflatées en zéro. Bien que beaucoup de méthodes d'inférence de réseaux existent pour les données gaussiennes ou les données binaires, les modèles inflatés en zéro sont très peu étudiés alors qu'ils reflètent la structure de nombreux jeux de données de façon pertinente. L'objectif de cette thèse concerne l'inférence de réseaux pour les modèles inflatés en zéro. Dans cette thèse, on se limitera à des graphes de dépendances conditionnelles. Le travail présenté dans cette thèse se décompose principalement en deux parties.

La première, composée des chapitres 3, 4 et 5, concerne des méthodes d'inférence de réseaux basées sur l'estimation de voisinages par une procédure couplant des méthodes de régressions ordinales et de sélection de variables. Cette façon d'inférer un réseau est similaire aux approches de Meinshausen et Bühlmann (2006) [71] dans le modèle graphique gaussien avec la régression linéaire, et de Wainwright *et al.* (2007) [100] dans le modèle d'Ising avec la régression logistique.

La seconde partie, que constitue le chapitre 6, se focalise sur l'inférence de réseaux dans un modèle dérivé du modèle graphique gaussien.

Chapitre 3 : Régression L_1-pénalisée à logits cumulatifs et odds proportionnels

Dans le chapitre 3, on travaille d'abord sur la régression pénalisée à logits cumulatifs (et odds proportionnels). Il s'agit d'une régression pour une variable réponse ordinale, qui tient compte du caractère ordonné des modalités de la variable réponse et permet ainsi de prendre en charge les 0 de celle-ci à part entière. Cette régression existait déjà dans la littérature ; sa version L_1-pénalisée n'existait en revanche pas quand nous avons entamé cette étude. Elle a entre temps été développée dans un package R. Ce chapitre traite de la version Lasso pénalisée de cette régression et de la résolution du problème d'optimisation associé que nous avions proposée. On s'intéresse en même temps à la sélection de variables, l'objectif étant, par la suite, de sélectionner un voisinage dans le cadre de l'inférence de réseaux. Des tests statistiques pour la sélection de

variables existent déjà dans la littérature pour la régression à logits cumulatifs. L'idée est de tester la nullité de chacun des coefficients de régression pour déterminer si la covariable associée appartient au modèle. Ces tests sont essentiellement basés sur la normalité asymptotique des estimateurs de ces coefficients ou sur le rapport de vraisemblance. Ce genre de procédure est cependant fastidieuse et chronophage, et devient, de plus, inappropriée quand le nombre de covariables excède le nombre d'observations. Le Lasso permet de surmonter ces problèmes, ce qui en fait actuellement une des méthodes les plus utilisées pour la sélection de variables. L'enjeu devient alors le calibrage du paramètre de pénalisation. Dans ce chapitre, nous proposons également une nouvelle méthode de sélection de variables, basée sur la méthode des knockoffs de Barber et Candès [7] que nous comparons à des méthodes existantes telles que la méthode classique de validation croisée ou la méthode de *stability selection* de Meinshausen et Bühlmann (2010) [72]. La validation croisée est une méthode conduisant à un choix du paramètre de pénalisation ; ce n'est en revanche pas le cas des deux autres méthodes.
Le travail de ce chapitre a été soumis dans un article en mai 2018, soumission que nous avons décidé d'abandonner après avoir découvert l'existence du package R qui proposait l'implémentation de la version L_1-pénalisée de cette régression.

Chapitre 4 : La méthode des knockoffs revisités pour la sélection de variables

Dans le chapitre 4, nous décrivons de façon plus détaillée la méthode des knockoffs revisités, introduite au chapitre 3. Cette méthode s'avère en réalité compatible avec un panel plus large de régressions, y compris quand le nombre d'observations est inférieur au nombre de covariables. Il s'agit d'une méthode relativement intuitive, qui utilise une matrice de copies (knockoffs) des covariables, non liées à la variable réponse mais dont la structure de corrélation est similaire à celle des covariables. L'idée est la même que celle de la méthode de Barber et Candès (2015) [7] : on va faire varier le paramètre de pénalisation et déterminer pour chacune des covariables si elle entre dans le modèle avant ou après sa copie. En d'autres termes, on va comparer les valeurs du paramètre de pénalisation pour lesquelles chacune des covariables et leur copie rentrent dans le modèle, c'est-à-dire pour quelle valeur du paramètre de pénalisation le coefficient de régression estimé associé devient non nul. Si une covariable entre dans le modèle après sa copie, qui est construite de façon indépendante de la variable réponse, on considère qu'elle n'est pas dans le modèle. Il reste ensuite à travailler sur les covariables restantes, c'est-à-dire celles qui sont entrées dans le modèle avant leur copie, et qui sont plus susceptibles d'appartenir au modèle.
Les deux principales différences entre notre procédure et la procédure originale résident dans la construction de la matrice des knockoffs et dans le choix du seuil final pour sélectionner les covariables parmi les covariables restantes (entrées dans le modèle avant leur copie). Dans notre cas, la matrice de knockoffs est construite par permutation aléatoire des lignes (qui correspondent aux observations) de la matrice des covariables et le seuil final est choisi grâce à des méthodes de détection de ruptures.
La méthode initiale de Barber et Candès est proposée dans le cadre de la régression linéaire gaussienne et convient quand le nombre d'observations est supérieur au nombre de covariables. La construction de leur matrice de knockoffs est plus sophistiquée et permet d'obtenir des garanties théoriques concernant le contrôle du *false discovery rate*, qui correspond à la proportion moyenne de faux positifs parmi les estimés positifs.
Notre procédure est implémentée dans le package R `kosel` disponible sur le CRAN.
Ce chapitre fait l'objet d'un article actuellement soumis à *Journal of Statistical Computation and Simulation.*

Chapitre 5 : Application de la méthode des knockoffs revisités pour la sélection de variables à l'inférence de réseaux pour données inflatées en zéro

Le chapitre 5 utilise les outils présentés dans les chapitres 3 et 4 pour les appliquer à l'inférence de réseaux dans des modèles inflatés en zéro. Plus précisément, nous allons construire le réseau estimé en estimant chacun des voisinages à l'aide de la régression pénalisée à logits cumulatifs ou d'une régression similaire, la régression adjacente, couplée à la méthode des knockoffs revisités. Nous appliquons cette procédure à deux modèles différents pour lesquels on connaît la structure sous-jacente de dépendances conditionnelles. Le premier est le produit d'un vecteur gaussien et d'un vecteur de variables de Bernoulli dépendant du vecteur gaussien, le vecteur gaussien donnant la structure de graphe et le vecteur de Bernoulli permettant l'inflation en zéro. Le second modèle est un modèle similaire au modèle d'Ising pour des variables à valeurs ordinales. Chacune des lois conditionnelles de ce modèle suit le modèle de régression adjacente évoqué ci-dessus, appelé aussi *adjacent-categories regression*, proche de la régression à logits cumulatifs présenté au chapitre 3. Nous avons choisi de privilégier ici ce modèle de régression adjacente, le modèle à logits cumulatifs n'étant pas compatible avec l'existence d'une loi jointe. En d'autres termes, il n'est en général pas possible de générer un vecteur tel que chacune de ses lois conditionnelles suive un modèle de régression à logits cumulatifs.
Nous appliquons ensuite aux données simulées selon ces deux modèles la procédure de régression à logits cumulatifs ou régression adjacente couplée à la méthode des knockoffs revisités pour l'estimation du voisinage de chacune des variables du graphe.
Nous présentons également une application à des données réelles, sur des données d'abondance de populations bactériennes se développant autour de la truffe.

Chapitre 6 : Inférence de réseaux pour données gaussiennes inflatées en zéro par double troncature

Le chapitre 6 se focalise sur l'inférence de réseaux dans un modèle dérivé du modèle graphique gaussien, où les variables sont des gaussiennes inflatées en zéro par double troncature (à droite et à gauche). La structure sous-jacente de graphe est donnée par les dépendances conditionnelles et donc par la matrice de précision du vecteur gaussien X. On note Y le vecteur inflaté en zéro, obtenu par double troncature du vecteur gaussien X. L'objectif est de retrouver cette structure de graphe à l'aide des observations inflatées en zéro par la troncature, c'est-à-dire les observations du vecteur Y. Pour ce faire, nous allons chercher à estimer la matrice de précision de X, dont les entrées non-nulles spécifient la structure de graphe sous-jacente.
Nous voudrions estimer cette matrice à l'aide de la procédure du graphical Lasso de Friedman *et al.* [36]. Cette procédure a été développée dans le modèle graphique gaussien et consiste à estimer la matrice de précision en maximisant une version L_1-pénalisée de la log-vraisemblance du modèle gaussien. Le problème est que cette procédure fait intervenir la matrice de covariance empirique des observations du vecteur gaussien et nous ne disposons ici que des observations inflatées en zéro par la troncature. Utiliser ces observations tronquées pour obtenir la matrice de covariance empirique du vecteur gaussien conduirait à une estimation pauvre. La première étape de notre procédure consiste donc à fournir une estimation plus adéquate de la matrice de covariance du vecteur gaussien (non observé directement). Notre estimateur de la matrice de covariance est obtenu en estimant chacune des entrées de la matrice à l'aide des vraisemblances des couples de variables du vecteur tronqué. Dans un second temps, on utilise la procédure du graphical Lasso dans laquelle on substitue la matrice de covariance empirique par notre estimateur obtenu à la première étape.

Nous obtenons alors deux principaux résultats théoriques. Le premier, donné par la proposition 6.2.2, est un résultat de convergence asymptotique de l'estimateur de la matrice de covariance obtenu à la première étape de notre procédure d'estimation. Cette proposition s'appuie sur les travaux de Mei *et al.* (2017) [68] et spécifie la vitesse de convergence en norme infinie. Le second, énoncé dans la proposition 6.2.3, est un résultat de convergence de l'estimateur de la matrice de précision en norme infinie, qui précise de plus des conditions sous lesquelles le graphe estimé coïncide avec le graphe théorique avec une forte probabilité. Cette proposition utilise les travaux menés par Ravikumar *et al.* (2011) [83] sur la procédure du graphical Lasso. La combinaison de ces deux résultats fournit le théorème 6.2.1 qui établit certaines garanties théoriques au sujet de la consistance de notre procédure quant à l'estimation du graphe théorique.
Un article concernant les résultats de ce chapitre est en préparation et sera soumis prochainement.

Deuxième partie

Régression pénalisée à logits cumulatifs et sélection de variables par la méthode des knockoffs

Dans cette partie, on s'intéresse à un modèle de régression pour estimer les voisinages dans le cadre d'inférence de réseaux de façon similaire à ce qu'ont fait Meinshausen et Bühlmann [71] pour le modèle gaussien ou Ravikumar *et al.* [82] pour le modèle d'Ising. Les données considérées pour l'inférence de réseaux sont inflatées en zéro, c'est-à-dire qu'elles comportent un fort nombre de zéros.

Comme on l'a vu précédemment, les méthodes de régression sont très utiles et largement répandues pour analyser les dépendances entre une variable réponse et des covariables [44]. Beaucoup de modèles de régression ont naturellement été introduits, comme la régression linéaire pour une variable réponse continue ou la régression logistique pour une variable réponse binaire. En effet, les données binaires sont omniprésentes dans plusieurs domaines comme la médecine (occurrence d'une maladie) ou l'économétrie (intentions de vote). D'autres types de données, comme les données catégorielles sont également très présentes. Parmi ce type de données, on distingue le cas ordonné (données ordinales) du cas non-ordonné (données nominales). On retrouve par exemple ce premier type de données dans les échelles de douleur, les différents stades d'un cancer, les notes sur les sites web ou encore dans les données collectées par des études (0 : très déçu, 10 : très satisfait). Plusieurs modèles ont été introduits pour des données ordinales : les *cumulative link models*, les *adjacent-categories models* ou encore les *continuation-ratio models* [66, 1, 63, 93]. Le choix d'un de ces modèles dépend du type de problème auquel on s'intéresse. Par exemple, les *continuation-ratio models* sont particulièrement adaptés pour des modèles de survie à travers le temps. Ici, on s'intéressera d'abord à un cas particulier des *cumulative link models* : la régression à logits cumulatifs avec odds proportionnels (*cumulative logit model with proportional odds*). Cette régression généralise la régression logistique pour une variable réponse discrète à modalités ordonnées à l'aide des probabilités cumulatives.

Même si la prédiction et l'interprétation constituent des motivations majeures de l'utilisation des régressions, un autre enjeu important est la sélection de variables. Notre but étant d'estimer les voisinages de chaque sommet pour l'inférence de réseaux, on s'intéresse en fait plus particulièrement à ce dernier enjeu, à savoir, identifier les covariables explicatives importantes qui ont une influence sur la variable réponse. On a déjà vu précédemment que ces problèmes de sélection survenaient dans plusieurs domaines, notamment la biologie où comprendre l'implication de certains gènes dans des maladies est primordial. Pour des raisons de coûts et de temps, il peut être commode pour les biologistes de restreindre leurs études à un plus petit ensemble de covariables (par exemple, de gènes). L'hypothèse de parcimonie, consistant à supposer que peu de covariables sont importantes (c'est-à-dire, impliquées dans le modèle), est souvent adéquate, voire cruciale pour l'interprétation. En effet, quand le nombre de covariables est colossal, l'interprétation devient également plus aisée en identifiant un sous-ensemble restreint de covariables qui sont les plus influentes. De plus, quand le nombre de covariables est plus grand que le nombre d'observations ou quand les covariables sont très corrélées, les méthodes de régressions usuelles deviennent inappropriées.
La pénalisation Lasso introduite par Tibshirani (1996) [94] offre une solution attractive à ces problèmes. Comme on l'a vu précédemment, elle consiste à introduire une pénalisation en norme L_1 dans l'estimation des coefficients, permettant de réduire à zéro les coefficients des covariables explicatives ayant peu voire pas d'effets sur la variable réponse. Cette pénalisation Lasso conduit alors à des solutions parcimonieuses et plus facilement interprétables, faisant d'elle une des pénalisations les plus populaires [119, 77, 45]. Cependant, l'introduction de cette pénalisation conduit souvent à de lourds problèmes d'optimisation. Dans notre cas, on résout le problème d'optimisation grâce à l'algorithme de Frank-Wolfe [35].

Utiliser une pénalisation Lasso implique également le choix délicat du paramètre de pénalisation qui contrôle le nombre de variables sélectionnées. Ce choix est primordial car deux valeurs proches du paramètre de pénalisation peuvent conduire à des solutions éloignées et donc à des conclusions scientifiques très différentes. De multiples techniques ont été proposées dans la littérature mais elles n'ont pas toujours les mêmes objectifs. Par exemple, la validation croisée met l'accent sur la prédiction, l'étape de validation visant à minimiser l'erreur de prédiction. De plus, la validation croisée est une méthode relativement gloutonne et a tendance à overfitter les données [103]. Hall (2009) *et al.* [42] et Wang *et al.* (2009) [102] ont par ailleurs montré qu'elle présentait de mauvaises performances pour le choix du paramètre avec le Lasso. D'autres techniques, comme StARS [61], peuvent être adaptées aux régressions et visent à "sursélectionner", c'est-à-dire à sélectionner un ensemble de covariables plus grand qui contient les covariables importantes quitte à sélectionner des faux positifs. Certains contextes comme les réseaux de régulation de gènes requièrent ce choix : en effet, les faux positifs peuvent être éliminés par des expérimentations biologiques supplémentaires tandis que des interactions oubliées ne peuvent être retrouvées après. On peut au contraire préférer sélectionner un ensemble de covariables inclus dans l'ensemble des "vraies" covariables importantes pour éviter les faux positifs ("sous-sélection"). Cette contrainte provient du fait qu'après la sélection, les covariables importantes doivent parfois faire l'objet de nouvelles expérimentations malheureusement coûteuses et chronophages.

On se concentre ici sur la seconde option. Pour répondre à cet objectif, nous avons développé une méthode de sélection de variables basée sur l'idée des knockoffs de Barber et Candès (2015) [7], qui fera l'objet du chapitre 4. Il s'agit d'une nouvelle méthode, intuitive et générale pour la sélection automatique de variables basée sur l'utilisation d'une matrice de copies des covariables pour déterminer si une covariable appartient au modèle. En outre, elle convient à un grand nombre de régressions, y compris quand le nombre d'observations n est plus petit que le nombre p de covariables. On présentera aussi une autre méthode existante compatible avec notre objectif : la méthode de *stability selection* introduite par Meinshausen et Bühlmann (2010) [72], qui consiste à estimer la probabilité qu'une covariable soit dans le modèle à l'aide de techniques de bootstrap. Contrairement à la validation croisée par exemple, aucune de ces deux méthodes ne fournit un choix du paramètre. Néanmoins, elles permettent de trier les covariables selon leur importance dans le modèle.

Dans le chapitre 3, on présentera d'abord en détails la régression L_1-pénalisée à logits cumulatifs, c'est-à-dire le modèle, l'estimation Lasso des coefficients et la résolution du problème d'optimisation qui en résulte. On présentera également des résultats de simulations obtenus avec les méthodes de sélection de variables mentionnées ci-dessus. Dans le chapitre 4, on présentera en détails notre méthode des knockoffs revisités pour la sélection de variables, qui s'adapte en fait à un spectre de régressions beaucoup plus large. Enfin, dans le chapitre 5, on s'intéressera plus spécifiquement à l'application de cette régression et de cette méthode de sélection de variables à l'inférence de réseaux dans le cadre de données inflatées en zéro.

Chapitre 3

Régression L_1-pénalisée à logits cumulatifs et odds proportionnels

Sommaire

3.1 Modèle à logits cumulatifs (et odds proportionnels) 43
3.1.1 Généralités 43
3.1.2 Interprétation des coefficients 46
3.2 Estimation et inférence 46
3.2.1 Estimation Lasso des coefficients β 47
3.2.2 Paramètre de pénalisation et sélection de variables 48
3.3 Simulations 52
3.3.1 Validation croisée 52
3.3.2 Distributions des covariables 53
3.3.3 Stability selection 54
3.3.4 Knockoffs revisités 58
3.4 Conclusions 65

Dans ce chapitre, on s'intéresse à la version Lasso pénalisée de la régression logistique à logits cumulatifs et odds proportionnels. Il s'agit d'une régression dont la variable réponse est ordinale, c'est-à-dire qu'elle prend un nombre fini de modalités ordonnées. Dans un premier temps, on présentera le modèle plus en détails, puis on s'intéressera à l'estimation Lasso des coefficients ainsi qu'aux problèmes qui en découlent : résolution du problème d'optimisation et gestion du paramètre de pénalisation Lasso pour la sélection de variables. Pour la sélection de variables, on présentera notamment quelques essais effectués avec la validation croisée et des simulations plus avancées obtenues avec notre méthode des knockoffs revisités qu'on comparera avec la méthode de *stability selection*.

3.1 Modèle à logits cumulatifs (et odds proportionnels)

3.1.1 Généralités

Le modèle à logits cumulatifs a été introduit et étudié par Williams et Grizzle (1972) [112], Simon (1974) [89], Agresti (1984 [1], 1990 [2]) ou encore McCullagh (1980) [66]. Il généralise la régression logistique pour une variable réponse Y qui prend $J > 2$ modalités ordonnées en utilisant les probabilités cumulatives.

Supposons qu'on ait p variables explicatives $(X_1, X_2, \ldots, X_p) =: X$ et notons $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1}) \in \mathbb{R}^{J-1}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ et $\boldsymbol{\beta}^* = (\alpha_1, \ldots, \alpha_{J-1}, \beta_1, \ldots, \beta_p) = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{J-1+p}$.

On modélise la probabilité cumulative $p^j_{\boldsymbol{\beta}^*}(x) := \mathbb{P}_{\boldsymbol{\beta}^*}(Y \leq j | X = x)$ pour $j = 1, \ldots, J-1$ et $x \in \mathbb{R}^p$, par :

$$\text{logit } p^j_{\boldsymbol{\beta}^*}(x) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{3.1}$$

i.e.,

$$p^j_{\boldsymbol{\beta}^*}(x) = \frac{\exp(\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p)}.$$

Le vecteur $\boldsymbol{\beta}$ des coefficients de régression ne dépend pas de la modalité j de la variable réponse. Ceci traduit un effet identique des covariables pour chaque probabilité cumulative. En fait, on peut supposer que ces coefficients β_k dépendent du niveau j pour permettre des effets séparés des covariables selon la modalité ; mais si x prend un spectre de valeurs trop étendu, ceci implique des courbes non parallèles pour différents logits et contredit alors l'ordre des probabilités cumulatives [2, 63]. Prenons un exemple basique $p = 1$ covariable et $J = 3$ modalités pour Y. En supposant que x prend ses valeurs dans un ensemble suffisamment étendu et que le coefficient β^j dépend de la modalité $j \in \{1, 2, 3\}$, on a : $\beta^1 \neq \beta^2$ et il existe alors des valeurs de x pour lesquelles $\alpha_1 + \beta^1 x > \alpha_2 + \beta^2 x$ (voir figure 3.1). La fonction logit^{-1} étant strictement croissante, on obtient alors pour ces valeurs de x que $\mathbb{P}(Y \leq 1 | X = x) > \mathbb{P}(Y \leq 2 | X = x)$.

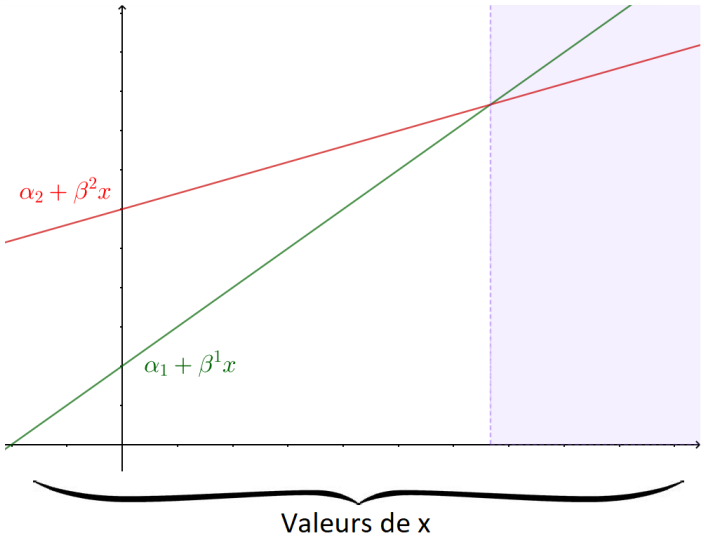


FIGURE 3.1 – Illustration du phénomène des courbes non parallèles quand les effets des covariables sont séparés (quand les coefficients de régression dépendent de la modalité). La zone violette correspond aux valeurs de x pour lesquelles l'ordre des probabilités cumulatives n'est pas respecté : $\mathbb{P}(Y \leq 1 | X = x) > \mathbb{P}(Y \leq 2 | X = x)$.

Bien que ce modèle d'effets séparés puisse être valide pour des covariables prenant leurs valeurs dans un ensemble restreint, il est préférable d'éviter ce modèle, en particulier lorsqu'on a peu d'informations sur les covariables. C'est pourquoi on se concentre sur le modèle plus simple d'effets similaires décrit en (3.1). Dans ce cas, on a alors une contrainte supplémentaire sur les

coefficients seuils α_j qui provient de l'ordre naturel des probabilités cumulatives :

$$\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}.$$

De la même façon que pour la régression logistique, ce modèle peut être expliqué par l'existence d'une variable latente continue [4] dont la distribution est logistique. En fait, si on introduit ϵ une variable aléatoire symétrique et une variable latente non observée : $Y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ telle que : $Y|X = x = \begin{cases} 1, \text{ si } Y^* < \tilde{\alpha}_1, \\ j, \text{ si } \tilde{\alpha}_{j-1} \leq Y^* < \tilde{\alpha}_j, j = 2, \ldots, J-1 \\ J, \text{ si } Y^* \geq \tilde{\alpha}_{J-1} \end{cases}$, où $\tilde{\alpha}_1 < \cdots < \tilde{\alpha}_{J-1}$ on obtient alors :

$$\mathbb{P}(Y \leq j|X = x) = \mathbb{P}(Y^* < \tilde{\alpha}_j) = \mathbb{P}(\epsilon < \tilde{\alpha}_j - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p) = F(\alpha_j - \beta_1 x_1 - \cdots - \beta_p x_p),$$

où F est la fonction de répartition de ϵ et $\alpha_j = \tilde{\alpha}_j - \beta_0$. La loi logistique pour ϵ conduit à notre modèle mais d'autres lois, c'est-à-dire d'autres fonctions de liens, sont envisageables. Dans la littérature [1, 63], on retrouve notamment les fonctions probit et log-log. La première correspond à l'inverse de la fonction de répartition de la loi $\mathcal{N}(0,1)$ (et donc à $\epsilon \sim \mathcal{N}(0,1)$), la seconde est la fonction : $p \mapsto \log(-\log(1-p)$. Les fonctions logit et probit ayant une forme similaire, les modèles associés conviennent généralement aux mêmes situations. Le modèle logit est cependant souvent privilégié car il est plus simple à manipuler et fournit une interprétation des coefficients en terme d'odds ratio (détaillée à la sous-section 3.1.2). Ces modèles sont tous inclus dans les plus généraux *cumulative link models* dont les *cumulative logit models* sont un cas particulier.

Comme $p^J_{\boldsymbol{\beta}^*}(x) = 1$, ce modèle comporte $p + J - 1$ coefficients à estimer : $J - 1$ coefficients seuils (le vecteur $\boldsymbol{\alpha}$) et p coefficients de régression (le vecteur $\boldsymbol{\beta}$). Supposons que $(Y^{(i)}, X_1^{(i)}, \ldots X_p^{(i)})_{1 \leq i \leq n}$ sont n vecteurs indépendants et identiquement distribués. On note $X_j^{*(i)} = (0, \ldots, 1, \ldots, 0, X_1^{(i)}, ..., X_p^{(i)})$ le $(J-1+p)$-vecteur où 1 est à la j^{eme} position, avec les conventions $\boldsymbol{\beta}^* X_0^{*(i)} = -\infty$ et $\boldsymbol{\beta}^* X_J^{*(i)} = +\infty$. On peut à présent donner une expression de :

— la log-vraisemblance :

$$L(\boldsymbol{\beta}^*) = \sum_{j=1}^{J} \sum_{i/Y^{(i)}=j} \log\Big[\frac{\exp(-\boldsymbol{\beta}^* X_{j-1}^{*(i)}) - \exp(-\boldsymbol{\beta}^* X_j^{*(i)})}{(1+\exp(-\boldsymbol{\beta}^* X_j^{*(i)}))(1+\exp(-\boldsymbol{\beta}^* X_{j-1}^{*(i)}))}\Big], \tag{3.2}$$

— le gradient de la log-vraisemblance :

$$\begin{aligned} \nabla L(\boldsymbol{\beta}^*) = \sum_{j=1}^{J} \sum_{i/Y^{(i)}=j} \Big[& \frac{X_j^{*(i)} \exp(\boldsymbol{\beta}^* X_j^{*(i)}) - X_{j-1}^{*(i)} \exp(\boldsymbol{\beta}^* X_{j-1}^{*(i)})}{\exp(\boldsymbol{\beta}^* X_j^{*(i)}) - \exp(\boldsymbol{\beta}^* X_{j-1}^{*(i)})} - \\ & \frac{X_j^{*(i)}}{1+\exp(-\boldsymbol{\beta}^* X_j^{*(i)})} - \frac{X_{j-1}^{*(i)}}{1+\exp(-\boldsymbol{\beta}^* X_{j-1}^{*(i)})}\Big]. \end{aligned} \tag{3.3}$$

Le gradient défini en (3.3) sera utile au moment de l'étape d'optimisation (voir sous-section 3.2.1).

3.1.2 Interprétation des coefficients

De la même façon que pour la régression logistique, on peut considérer les odds et odds ratios pour le modèle à logits cumulatifs (à effets parallèles).

Pour $X = x$ fixé et pour tout $j \in \{1, \dots, J-1\}$, les odds sont définis par :

$$\mathrm{odd}_j(x) = \exp(\mathrm{logit}(p^j_{\boldsymbol{\beta}^*}(x))) = \frac{\mathbb{P}(Y \leq j | X = x)}{1 - \mathbb{P}(Y \leq j | X = x)} = \exp(\alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p).$$

Ce rapport mesure la tendance de Y à être plus grand ou plus petit que j sachant $X = x$. Si l'odd est grand : Y a plutôt tendance à être plus petit que j sachant $X = x$.

On considère également les odds ratios, c'est-à-dire les rapports d'odds des probabilités cumulatives qui sont définis par : $\dfrac{\mathrm{odd}_j(x)}{\mathrm{odd}_j(\tilde{x})}$ pour tout $x, \tilde{x} \in \mathbb{R}^p$. Si ce rapport est grand, Y a tendance à être plus petit que j sous $X = x$ que sous $X = \tilde{x}$. Dans le cas particulier où $\tilde{x} = x_k^{+z} := (x_1, \dots, x_k + z, \dots, x_p)$, alors pour tout $j \in \{1, \dots, J-1\}$, $\dfrac{\mathrm{odd}_j(x)}{\mathrm{odd}_j(x_k^{+z})} = \exp(-\beta_k z)$.

On remarque que ces odds ratios cumulatifs sont les mêmes pour n'importe quel niveau j. Ceci provient de l'hypothèse sur les effets identiques des covariables et ne serait plus vrai si les coefficients de régression β_k dépendaient de la modalité j (quand les effets des covariables ne sont plus identiques). Dans la littérature, ce modèle est de ce fait parfois appelé "régression logistique à logits cumulatifs et odds proportionnels" ou "régression logistique à logits cumulatifs et effets parallèles" [66].

Certaines références privilégient la paramétrisation logit $p^j_{\boldsymbol{\beta}^*}(x) = \alpha_j - \beta_1 x_1 - \cdots - \beta_p x_p$, provenant de l'explication du modèle par une variable latente, au lieu de celle des "+" utilisée en (3.1). Ce choix de paramétrisation affecte uniquement l'interprétation des odds ratios, le signe "−" correspondant à l'interprétation habituelle [63] en terme d'effet "positif" ou "négatif" de la covariable relative.

Notre objectif est de sélectionner les variables importantes, à savoir les covariables X_k dont le coefficient de régression β_k est non-nul. On remarque que β_k mesure également la dépendance conditionnelle entre Y et X_k sachant $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_p$. C'est pourquoi on s'intéresse particulièrement à la nullité de ces coefficients de régression β_k. On fait ici de plus une hypothèse de parcimonie, c'est-à-dire qu'on suppose que seulement peu de β_k sont non nuls, ce qui correspond à supposer que seul un petit nombre de covariables jouent un rôle important et sont impliquées dans le modèle. Cette hypothèse de parcimonie est pratique pour les scientifiques afin de restreindre leurs études à un sous-ensemble plus petit de covariables, notamment dans un contexte de grande dimension. Au lieu de vérifier la nullité de chacun des coefficients de régression β_k à l'aide de tests statistiques, on va ajouter une pénalisation en norme L_1 sur le vecteur $\boldsymbol{\beta}$ de coefficients de régression dans la log-vraisemblance lors de l'estimation.

3.2 Estimation et inférence

Sans l'hypothèse de parcimonie, les coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ sont estimés par maximisation de la log-vraisemblance L à l'aide d'algorithmes de scoring de Fisher [101, 66].

3.2.1 Estimation Lasso des coefficients $\boldsymbol{\beta}$

Pour garantir l'hypothèse de parcimonie, on pénalise la log-vraisemblance L sur le vecteur $\boldsymbol{\beta}$ des coefficients de régression. Pour cela, on est donc ramené à résoudre le problème d'optimisation suivant :

$$\underset{\substack{\boldsymbol{\alpha}\in A\\ \boldsymbol{\beta}\in\mathbb{R}^p}}{\operatorname{argmax}} \{L(\boldsymbol{\alpha},\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1\}, \tag{3.4}$$

où A est l'ensemble convexe de $\mathbb{R}^{J-1}$ défini par : $A := \{(\alpha_1,...,\alpha_{J-1}) \in \mathbb{R}^{J-1} \,/\, \alpha_1 < \ldots < \alpha_{J-1}\}$, $||.||_1$ correspond à la norme L_1, c'est-à-dire $||\boldsymbol{\beta}||_1 = \sum_{i=1}^{p} |\beta_i|$ et $\lambda > 0$ est le paramètre de pénalisation. Par dualité lagrangienne, résoudre le problème d'optimisation (3.4) est équivalent à résoudre le problème :

$$\underset{\substack{\boldsymbol{\alpha}\in A\\ \boldsymbol{\beta}\in B_\tau}}{\operatorname{argmax}} L(\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{3.5}$$

où B_τ est l'ensemble convexe suivant de $\mathbb{R}^p$: $B_\tau := \{(\beta_1,...,\beta_p) \in \mathbb{R}^p \,/\, ||\boldsymbol{\beta}||_1 \leq \tau\}$ et $\tau > 0$ est alors le "nouveau" paramètre de pénalisation. Il y a une correspondance bijective implicite entre les paramètres de pénalisation λ et τ [45].

On propose de résoudre ce problème d'optimisation grâce à l'algorithme de Frank-Wolfe [35] décrit plus en détails ci-dessous. L'idée est de remplacer la fonction cible à optimiser par une approximation linéaire. Dans notre cas, on utilise le gradient donné en (3.3) pour approximer la log-vraisemblance L donnée en (3.2).

Algorithme de Frank-Wolfe

– Étape 1 : Commencer avec une valeur initiale $\boldsymbol{\beta}^*_{\mathbf{0}} = (\boldsymbol{\alpha_0}, \boldsymbol{\beta_0})$.

– Étape 2 : À chaque itération k,

- Résoudre $s_k = (s_{k,\boldsymbol{\alpha}}, s_{k,\boldsymbol{\beta}}) \in \underset{\substack{s_{\boldsymbol{\alpha}}\in A\\ s_{\boldsymbol{\beta}}\in B_\tau}}{\operatorname{argmin}} (-\nabla L(\boldsymbol{\beta}^*_{\boldsymbol{k}}))' \begin{pmatrix} s_{\boldsymbol{\alpha}} \\ s_{\boldsymbol{\beta}} \end{pmatrix}$.
- Le nouvel itéré est $\boldsymbol{\beta}^*_{\boldsymbol{k+1}} = (1-\gamma_k)\boldsymbol{\beta}^*_{\boldsymbol{k}} + \gamma_k s_k$, où $\gamma_k = \frac{2}{k+1}$.

– Étape 3 : Itérer l'étape 2 jusqu'à convergence.

L'étape 2 utilise la convexité des ensembles A et B_τ, le nouvel itéré étant une combinaison convexe d'éléments de ces ensembles. On peut séparer le problème d'optimisation de l'étape 2 en deux problèmes d'optimisation :

$$\underset{\substack{s_{\boldsymbol{\alpha}}\in A\\ s_{\boldsymbol{\beta}}\in B_\tau}}{\operatorname{argmin}} (-\nabla L(\boldsymbol{\beta}^*_{\boldsymbol{k}}))' \begin{pmatrix} s_{\boldsymbol{\alpha}} \\ s_{\boldsymbol{\beta}} \end{pmatrix} = \underset{\substack{s_{\boldsymbol{\alpha}}\in A\\ s_{\boldsymbol{\beta}}\in B_\tau}}{\operatorname{argmin}} \Big((-\nabla L(\boldsymbol{\beta}^*_{\boldsymbol{k}}))'_{|\boldsymbol{\alpha}} s_{\boldsymbol{\alpha}} + (-\nabla L(\boldsymbol{\beta}^*_{\boldsymbol{k}}))'_{|\boldsymbol{\beta}} s_{\boldsymbol{\beta}} \Big).$$

Le premier problème d'optimisation concerne le vecteur $\boldsymbol{\alpha}$ des coefficients seuils :

$$s_{\boldsymbol{\alpha}} \in \underset{s\in A}{\operatorname{argmin}} (-\nabla L(\boldsymbol{\beta}^*_{\boldsymbol{k}}))'_{|\boldsymbol{\alpha}} s,$$

et s'avère être un problème d'optimisation linéaire sous contraintes. Le second concerne le vecteur $\boldsymbol{\beta}$ des coefficients de régression :

$$s_{\boldsymbol{\beta}} \in \underset{s \in B_\tau}{\operatorname{argmin}} \ (-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*))'_{|\boldsymbol{\beta}} s. \tag{3.6}$$

La problème d'optimisation (3.6) relatif au vecteur $\boldsymbol{\beta}$ est équivalent à résoudre :

$$-\tau \ \underset{s \in B_1}{\operatorname{argmax}} \ (-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*))'_{|\boldsymbol{\beta}} s. \tag{3.7}$$

On doit donc résoudre un problème du type : $\underset{||s||_1 \leq 1}{\operatorname{argmax}} \sum_{i=1}^{p} v_i s_i$ où $v \in \mathbb{R}^p$. Or $\sum_{i=1}^{p} v_i s_i \leq \max_i |v_i|$ pour tout $s \in B_1 = \{s \in \mathbb{R}^p : ||s||_1 \leq 1\}$. La résolution de ce problème consiste donc à choisir un vecteur e_{i_k} de la base canonique $(e_1, \ldots, e_p)$ tel que $|v_{i_k}| = ||v||_\infty := \max_i |v_i|$. Une solution est alors $\text{sign}(v_{i_k}) e_{i_k}$.
Appliqué à notre problème (3.7), ceci revient à choisir une coordonnée $i_k \in \{1, \ldots, p\}$ telle que $|(-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*)_{|\boldsymbol{\beta}})_{i_k}|$ maximise la norme infinie du gradient (restreinte au vecteur $\boldsymbol{\beta}$), c'est-à-dire $|(-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*)_{|\boldsymbol{\beta}})_{i_k}| = \max_{i=1,\ldots,p} |(-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*)_{|\boldsymbol{\beta}})_i|$. Cette coordonnée i_k n'est pas nécessairement unique.
Une solution du problème d'optimisation (3.6) est alors :

$$s_{\boldsymbol{\beta}} = -\tau \ \text{sign}(-\nabla L(\boldsymbol{\beta}_{\boldsymbol{k}}^*)_{|\boldsymbol{\beta}})_{i_k} \ e_{i_k},$$

où e_i est le i^{eme} vecteur de la base canonique.

3.2.2 Paramètre de pénalisation et sélection de variables

Introduire une pénalisation requiert le choix du paramètre de pénalisation, parfois aussi appelé paramètre de régularisation ou paramètre de réglage. On a donc besoin ici de régler ce paramètre de pénalisation τ qui apparaît dans la contrainte du problème d'optimisation (3.5). Ce paramètre contrôle le nombre de variables sélectionnées : si $\tau = 0$, aucune variable n'est sélectionnée et le modèle estimé est vide (il ne contient aucune covariable). Plus τ augmente, plus le nombre de variables sélectionnées augmente. On rappelle que notre objectif est de sélectionner seulement des variables appartenant au modèle. On cherche donc à limiter le nombre de faux positifs, c'est-à-dire le nombre de covariables détectées à tort dans le modèle estimé.

On s'est penché pour cela sur plusieurs méthodes. Dans un premier temps, on a essayé de choisir le paramètre de pénalisation à l'aide de la méthode de validation croisée, qui, en plus d'être relativement chronophage, s'est avérée être peu efficace au regard de notre but. On s'est ensuite plutôt attardé sur deux méthodes davantage compatibles avec notre objectif. La première est la méthode de *stability selection* proposée par Meinshausen et Bühlmann [72] et la deuxième est une méthode que nous avons développée, inspirée de la méthode des knockoffs de Barber et Candès [7], qui sera exposée en détails dans le chapitre 4. En fait, ces deux méthodes ne fournissent pas un choix du paramètre de pénalisation τ mais elles classent les covariables en fonction de leur importance dans le modèle. Par ailleurs, elles conviennent toutes deux à plusieurs types de régression, y compris dans la configuration où le nombre d'observations n est plus petit que le nombre p de variables explicatives. On présente ici le principe de ces trois méthodes avant de donner quelques exemples et résultats de simulations dans la section 3.3.

Validation croisée

Le principe de la validation croisée est de partitionner l'échantillon en M sous-échantillons $E_1, E_2, ..., E_M$ d'environ même taille. On fixe une valeur du paramètre de pénalisation τ et pour chaque $m = 1, \ldots, M$, on estime les coefficients $\beta_\tau^{(-m)}$ sur l'échantillon total privé du sous-échantillon E_m, c'est-à-dire sur les $M-1$ sous-échantillons $E_i,\ i \neq m$. On calcule ensuite l'erreur (de prédiction) commise sur l'échantillon E_m notée $err_m(\tau)$ (étape de validation). À τ fixé, ceci fournit alors M valeurs de l'erreur commise : $err_m(\tau),\ m = 1, \ldots, M$. On peut ainsi calculer l'erreur moyenne pour ce paramètre τ : $err(\tau) = \frac{1}{M} \sum_{m=1}^{M} err_m(\tau)$. On choisit ensuite la valeur du paramètre τ qui minimise la fonction err :

$$\tau_{CV} := \underset{\tau>0}{\operatorname{argmin}}\ err(\tau).$$

Pour la régression linéaire classique, la fonction d'erreur est :

$$err_m(\lambda) = \frac{1}{|E_m|} \sum_{i \in E_m} \big(Y^{(i)} - (x^{(i)})^T \hat{\beta}_\lambda^{(-m)}\big)^2,$$

où $\lambda > 0$ est le paramètre de pénalisation et $\beta_\lambda^{(-m)}$ désigne les coefficients de régression β estimés sur les sous-échantillons $E_i,\ i \neq m$ pour le paramètre de pénalisation λ. Pour la régression logistique, pour $m \in \{1, \ldots, M\}$ fixé, on calcule $p_{\hat{\beta}_\lambda^{(-m)}}(x^{(i)})$ et on prédit $\hat{Y}^{(i)}$ pour tout $i \in E_m$: la plupart utilisent le seuil 0.5 c'est-à-dire que si $p_{\hat{\beta}_\lambda^{(-m)}}(x^{(i)}) > 0.5$, alors on prédit $\hat{Y}^{(i)} = 1$. La fonction d'erreur est alors :

$$\begin{aligned} err_m(\lambda) &= \frac{1}{|E_m|} \sum_{i \in E_m} \big(Y^{(i)} - \hat{Y}^{(i)}\big)^2 \\ &= \frac{1}{|E_m|} \sum_{i \in E_m} |Y^{(i)} - \hat{Y}^{(i)}|. \end{aligned}$$

La validation croisée est une méthode répandue qui met l'accent sur la prédiction. En effet, le paramètre de pénalisation choisi à l'issue de la procédure est le paramètre qui minimise une fonction d'erreur basée sur la prédiction. De plus, cette méthode est assez gourmande et a tendance à overfitter les données [103]. Par ailleurs, [42] et [102] ont montré qu'elle présentait de mauvaises performances pour le choix du paramètre avec le Lasso. Par curiosité et parce qu'il s'agit tout de même d'une méthode très classique de sélection de variables, on a voulu l'essayer pour notre modèle. Pour cela, il s'agit de déterminer des fonctions d'erreur qui conviennent à la régression logistique à logits cumulatifs.

De la même façon que pour la régression logistique, on a besoin d'une prédiction pour la variable réponse. Pour une valeur fixée du paramètre de pénalisation τ et pour $m \in \{1, \ldots, M\}$ fixé, on calcule les $J-1$ probabilités cumulatives estimées :

$$\begin{aligned} p^j_{\hat{\boldsymbol{\beta}}_\tau^{(-m)}}(x^{(i)}) &= \mathbb{P}_{\hat{\boldsymbol{\beta}}_\tau^{(-m)}}(Y \leq j | X = x^{(i)}) \\ &= \text{logit}^{-1}(\hat{\alpha}_{\tau,j}^{(-m)} + \hat{\beta}_{\tau,1}^{(-m)} x_1^{(i)} + \cdots + \hat{\beta}_{\tau,p}^{(-m)} x_p^{(i)}), \end{aligned}$$

pour $1 \leq j < J$ et pour chacune des observations $x^{(i)}$ du sous-échantillon E_m. À partir de ces probabilités estimées, on peut alors établir une règle de prédiction de la variable réponse Y. On choisit la modalité dont la probabilité estimée est la plus élevée. On obtient alors la prédiction $\hat{Y}^{(i)}$ correspondante et on considère les différentes fonctions d'erreur :

— $ErrDist_m(\tau) = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} \mathbb{1}_{Y^{(i)} \neq \hat{Y}^{(i)}}$. Cette fonction d'erreur revient à compter combien de fois on s'est trompé dans la prédiction et ne tient pas compte du caractère ordonné des modalités de Y.

— $ErrDistAbs_m(\tau) = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} |Y^{(i)} - \hat{Y}^{(i)}|$. Dans ce cas, on pénalise davantage si la prédiction est loin de la valeur initiale.

— $ErrPropEG_m(\tau) = \sum\limits_{j=1}^{J} ErrPropEG_m^j(\tau) = \sum\limits_{j=1}^{J} \dfrac{(th_j^m - \theta_j^m)^2}{th_j^m}$ où pour tout $j \in \{1, \ldots, J\}$:

- $th_j^m = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} \left[p^j_{\hat{\boldsymbol{\beta}}_\tau^{(-m)}}(x^{(i)}) - p^{j-1}_{\hat{\boldsymbol{\beta}}_\tau^{(-m)}}(x^{(i)}) \right]$, où la quantité entre crochets correspond à la probabilité estimée d'être égal à j sachant $X = x^{(i)}$.
- $\theta_j^m = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} \mathbb{1}_{Y^{(i)}=j}$, la proportion de Y égaux à j dans le sous-échantillon E_m.

Cette fonction d'erreur ne tient pas non plus compte du caractère ordonné des modalités.

— La dernière fonction est une version modifiée de la précédente pour tenir compte de l'ordre : $ErrProp_m(\tau) = \sum\limits_{j=1}^{J-1} ErrProp_m^j(\tau) = \sum\limits_{j=1}^{J-1} \dfrac{(th_j^m - \theta_j^m)^2}{th_j^m}$ où pour tout $j \in \{1, \ldots, J-1\}$:

- $th_j^m = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} \left[p^j_{\hat{\boldsymbol{\beta}}_\tau^{(-m)}}(x^{(i)}) \right]$, où la quantité entre crochets correspond à la probabilité estimée d'être inférieur ou égal à j sachant $X = x^{(i)}$.
- $\theta_j^m = \dfrac{1}{|E_m|} \sum\limits_{i \in E_m} \mathbb{1}_{Y^{(i)} \leq j}$, la proportion de Y inférieurs ou égaux à j dans le sous-échantillon E_m.

Cette fonction d'erreur tient compte de l'ordre des modalités mais elle va avoir tendance à pénaliser davantage les erreurs commises sur les petites modalités de la variable réponse (qui sont comptées plusieurs fois).

Les résultats de simulations concernant les essais qui ont été faits avec ces fonctions d'erreur sont présentés à la sous-section 3.3.1.

Stability selection

Le principe de cette méthode est d'estimer la probabilité qu'une covariable appartienne au modèle afin de déterminer quelles covariables sont les plus importantes.

On considère un ensemble T de valeurs pour le paramètre de pénalisation τ. Pour chaque $\tau \in T$, l'idée est d'estimer la probabilité $p_k(\tau)$ que la covariable X_k appartienne au modèle. Pour cela, on effectue la régression pénalisée sur B ensembles de n observations obtenus par bootstrap. La probabilité estimée $\hat{p}_k(\tau)$ est alors la proportion de sélection de la covariable X_k parmi les B régressions bootstrappées pour cette valeur fixée de τ.

Généralement, la sélection de variables consiste à choisir un paramètre de pénalisation τ. Ceci revient quelque part à choisir un modèle estimé parmi $\{\hat{S}^\tau,\ \tau \in T\}$ où $\hat{S}^\tau$ est le modèle estimé relatif au paramètre fixé τ :

$$\hat{S}^\tau = \{X_k : k \in \{1, ..., p\} \text{ tels que } \hat{\beta}_k(\tau) \neq 0\},$$

où $\hat{\boldsymbol{\beta}}(\tau)$ désigne les coefficients estimés de la régression τ-penalisée de Y sur les covariables $X_1, \ldots, X_p$. Au lieu de choisir un de ces modèles, on choisit ici le modèle :

$$\hat{S} := \{X_k : k \in \{1, ..., p\} \text{ tels que } \max_{\tau \in T} \hat{p}_k(\tau) \geq p_{thr}\}$$

pour un seuil fixé p_{thr} (voir [72] pour plus de détails). Les probabilités estimées $\max_{\tau \in T} \hat{p}_k(\tau)$ fournissent naturellement un ordre d'importance des covariables correspondantes dans le modèle.

La valeur du seuil p_{thr} a une faible influence, ce qui est très commode en comparaison de la sensibilité du paramètre de pénalisation τ. En effet, les résultats ont tendance à être très similaires pour des valeurs de p_{thr} relativement étendues, contrairement au paramètre de pénalisation τ pour lequel des valeurs proches peuvent donner des résultats très différents. Cependant, Meinshausen et Bühlmann [72] proposent tout de même une procédure pour choisir ce seuil p_{thr} ainsi que la région de régularisation T. Sous certaines hypothèses simplificatrices, leur procédure fournit une majoration du nombre moyen de faux positifs. Néanmoins, cette borne dépend d'une quantité inconnue, dépendant de T, et les hypothèses sont relativement fortes. C'est pourquoi on préfère utiliser un seuil différent, suggéré également par ces auteurs pour une utilisation en pratique et qui sera détaillé dans la sous-section 3.3.3.

Knockoffs revisités

On va ici donner très brièvement le principe de notre méthode des knockoffs revisités, qui fera l'objet du chapitre 4.

Par abus de notation, on notera ici X la matrice $n \times p$ des n observations du vecteur $(X_1, \ldots, X_p)$, appelée la matrice design. Le principe, donné par Barber et Candès [7] et développé dans la section 4.1 du chapitre suivant, est d'utiliser une matrice $\tilde{X}$ de copies des covariables X_k dont la structure de covariance est similaire à celle de X mais indépendante de Y. On augmente progressivement le paramètre de pénalisation τ de 0 à $+\infty$ et on va déterminer la plus petite valeur de τ pour laquelle chaque covariable et chaque copie entre dans le modèle, c'est-à-dire la plus petite valeur de τ pour laquelle le coefficient de régression estimé associé est non-nul. Le but est alors de déterminer si une covariable X_k appartient au modèle en étudiant si elle entre dans le modèle avant ou après sa copie $\tilde{X}_k$, c'est-à-dire pour une valeur de τ inférieure ou supérieure à celle de sa copie. En effet, comme chacune des copies est construite de manière à être indépendante de la variable réponse Y, si une covariable rentre dans le modèle après sa copie, on peut légitimement suspecter que cette covariable n'appartient pas au modèle. En revanche, si elle rentre dans le modèle avant sa copie, elle est plus susceptible d'appartenir au modèle mais cela n'est pas nécessairement le cas.

Les principales différences de notre méthode avec celle de Barber et Candès résident dans la construction de la matrice des knockoffs $\tilde{X}$ et dans la façon de déterminer les covariables appartenant au modèle parmi celles entrées avant leur copie.

Cette procédure se découpe donc en deux étapes :

— une première étape qui consiste à construire des statistiques qui permettent de mesurer si une covariable rentre dans le modèle avant ou après sa copie. Cette première étape permet de trier les covariables selon leur importance dans le modèle : les covariables qui rentrent dans le modèle après leur copie sont les moins importantes (car les moins susceptibles d'appartenir au modèle) ; celles qui rentrent avant leur copie sont les plus importantes. Parmi cette deuxième catégorie, on peut encore affiner le tri en fonction de la valeur du paramètre de pénalisation pour lequel ces covariables sont entrées dans le modèle.

— une deuxième étape qui consiste à choisir un seuil, basé sur des méthodes de détection de ruptures, pour déterminer les covariables appartenant au modèle parmi celles qui sont entrées avant leur copie.

3.3 Simulations

Cette section contient des résultats expérimentaux concernant l'optimisation Lasso réalisée avec l'algorithme de Frank-Wolfe et des comparaisons des différentes méthodes de sélection de variables présentées à la sous-section 3.2.2. On souhaite dans un premier temps exposer les essais qui ont été réalisés concernant la validation croisée puis dans un second temps, présenter les résultats concernant les deux autres méthodes. Notre but est d'une part, illustrer ce que produit la méthode de *stability selection* et d'autre part, étudier l'efficacité de notre méthode des knockoffs revisités.

3.3.1 Validation croisée

On présente ici les essais qui ont été faits pour le choix du paramètre de pénalisation à l'aide de la méthode de validation croisée. Pour avoir une idée de la performance des fonctions d'erreur présentées à la sous-section 3.2.2, on a effectué des simulations sur des petits jeux de données. Pour cela, on a simulé $n = 200$ observations de $p = 20$ covariables gaussiennes centrées réduites et indépendantes. Y prend $J = 3$ modalités et les coefficients valent $\boldsymbol{\alpha} = (-3, 3)$ et $\boldsymbol{\beta} = (-4, 3, 2, -1, 0, \ldots, 0)$. Ainsi, seules les quatre premières covariables sont dans le modèle. L'ensemble T dans lequel varie le paramètre de pénalisation τ est $T = \{0.1, 0.3, 0.5, \ldots, 10.1\}$. La figure 3.2 représente le nombre de détection de chacune des covariables pour les fonctions $ErrDist$, $ErrDistAbs$, $ErrPropEG$ et $ErrProp$ sur 100 répétitions.

Résultats et commentaires. Aucune de ces fonctions ne donne de résultats véritablement satisfaisants par rapport à notre objectif de "sous-sélection".
La fonction $ErrPropEG$ a tendance à mener à un choix de paramètre de pénalisation τ trop petit qui ne sélectionne quasiment que la première variable (dont le coefficient de régression est le plus fort). Les autres fonctions ont plutôt tendance à choisir un paramètre de pénalisation trop grand qui sélectionne aussi trop souvent les covariables nulles (environ une fois sur deux), c'est-à-dire les covariables dont le coefficient de régression est nul et qui ne sont donc pas dans le modèle. Finalement, il y a peu de différence entre les fonctions $ErrDist$ et $ErrDistAbs$. Ceci est toutefois peut-être dû au fait que la variable réponse Y ne prend que 3 modalités. Ainsi, l'écart entre les valeurs $|Y^{(i)} - \hat{Y}^{(i)}|$ ne pénalise pas beaucoup plus que $\mathbb{1}_{Y^{(i)} \neq \hat{Y}^{(i)}}$ (on pénalise de 2 au lieu de 1 au maximum).

Au vu de ces résultats et compte-tenu des résultats théoriques existants de la littérature, on a fait le choix de ne pas s'attarder sur cette méthode de validation croisée et de se concentrer

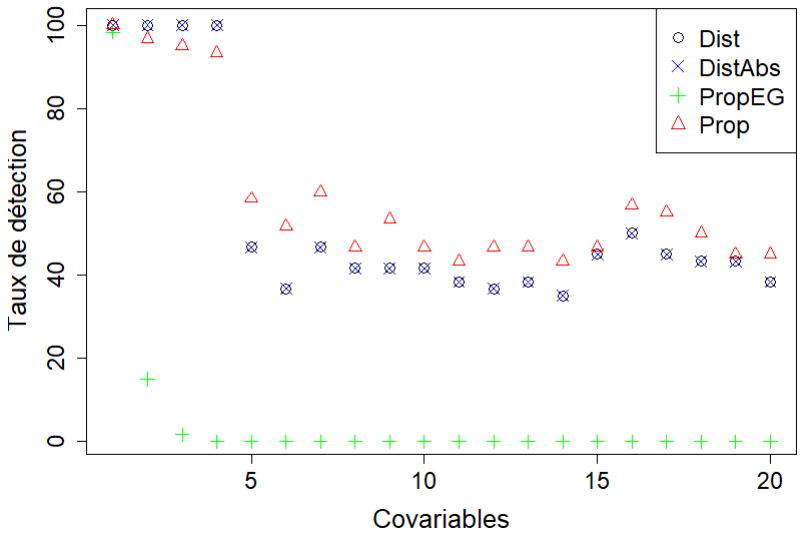


FIGURE 3.2 – Taux de détection de chacune des covariables avec le paramètre de pénalisation τ choisi par validation croisée pour les différentes fonctions d'erreur : $ErrDist$, $ErrDistAbs$, $ErrPropEG$ et $ErrProp$ introduites à la sous-section 3.2.2. Les coefficients de régression valent $\beta = (-4, 3, 2, -1, 0, \ldots, 0)$. Les taux de détection sont obtenus sur 100 répétitions de $n = 200$ observations de $p = 20$ covariables gaussiennes centrées réduites et indépendantes.

sur les méthodes de *stability selection* et knockoffs revisités.

3.3.2 Distributions des covariables

On présente maintenant les résultats concernant les deux autres méthodes (*stability selection* et les knockoffs revisités). Différentes distributions ont été utilisées pour simuler les covariables X :

Loi 1 : Les covariables X sont gaussiennes et indépendantes : $X \sim \mathcal{N}_p(0, I_p)$.

Loi 2 : Les covariables X sont indépendantes et simulées selon un mélange de lois comme suit :

- si $i \equiv 1 \pmod 3$, $X_i \sim \mathcal{N}(0, 1)$
- si $i \equiv 2 \pmod 3$, $X_i = \dfrac{Z_i - \mu}{\mu}$ où $Z_i \sim \mathcal{P}(\mu)$ et μ choisi uniformément sur $\{1, \ldots, 40\}$.
- si $i \equiv 0 \pmod 3$, $X_i \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$.

Les paramètres de ces lois sont choisis de sorte que les variables soient centrées et réduites.

Loi 3 : Les covariables X sont gaussiennes et X_j et X_k sont indépendantes conditionnellement aux autres variables avec probabilité 0.4. Le vecteur X des covariables est simulé à l'aide de la fonction R `huge.generator` issue du package `huge`, avec une structure "random graph". La plupart des corrélations non nulles sont entre -0.3 et 0.3 et les corrélations partielles non nulles valent environ -0.13.

3.3.3 Stability selection

Dans ce qui suit, on présente des résultats de simulations pour $n = 100$ et $n = 200$ observations de $p = 50$ covariables et $J = 3$ modalités pour la variable réponse Y. Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \dots, 0)$, ainsi, seules les quatre premières covariables appartiennent au modèle. Les seuils $\boldsymbol{\alpha}$ sont choisis de façon à ce que la variable réponse Y prenne assez de valeurs dans chacune de ses trois modalités. $p = 50$ covariables est un nombre relativement faible. Ces résultats de simulations sont en fait donnés à titre illustratif, la méthode de *stability selection* étant relativement coûteuse en raison des bootstraps et ne représentant par ailleurs pas le cœur de ce travail.

Paramètres de simulations. On a choisi d'estimer les probabilités sur $B = 100$ échantillons bootstrap et l'ensemble T des valeurs pour le paramètre de pénalisation τ est $T = \{0.1, 0.4, 0.7, \dots, \tau_{max}\}$. La valeur maximale τ_{max} de cet ensemble T a été choisie selon une suggestion de [72], il s'agit de la plus petite valeur telle qu'au moins $\sqrt{0.8p}$ covariables soient rentrées dans le modèle dans le chemin de solution pour $\tau \leq \tau_{max}$:

$$\tau_{max} = \min\left\{ t \in 0.1 + 0.3\mathbb{N} : \#\{k : \exists s \in 0.1 + 0.3\mathbb{N}, s \leq t,\ \hat{p}_k(s) \neq 0\} \geq \sqrt{0.8p} \right\}.$$

L'ensemble T ainsi que τ_{max} peuvent bien évidemment être modifiés mais on souligne tout de même que le choix du τ_{max} est d'autant plus délicat que de trop grandes valeurs du paramètre de pénalisation τ conduisent au modèle complet (pour tout k, $\hat{p}_k(\tau) = 1$). Le seuil p_{thr} peut également être modifié selon les besoins.

On représente les résultats par des boxplots (figures 3.3, 3.4 et 3.5) et des courbes ROC (figure 3.6). Les boxplots sont tous obtenus sur 50 répétitions de $n = 100$ ou 200 observations de $p = 50$ covariables et représentent le maximum des probabilités estimées $\max_{\tau \in T} \hat{p}_k(\tau)$ pour chacune des covariables X_k, $k = 1, \dots, p$. Les courbes ROC représentent le *true positive rate* (TPR) moyen en fonction du *false positive rate* (FPR) moyen obtenus après avoir seuillé pour $p_{thr} \in \{0.1, 0.15, 0.2, \dots, 1\}$ sur 50 répétitions. On rappelle que le *true positive rate* (ou sensibilité) vaut $\text{TPR} = \dfrac{TP}{TP + FN}$ et correspond à la proportion de vrais positifs parmi les positifs théoriques. Le *false positive rate* FPR vaut $\text{FPR} = \dfrac{FP}{TN + FP} = 1-$ Spécificité correspond, quant à lui, à la proportion de faux positifs parmi les négatifs théoriques. L'objectif étant de se rapprocher d'une estimation parfaite, on cherche alors à limiter à la fois les faux positifs et les faux négatifs, c'est-à-dire d'avoir un TPR proche de 1 et un FPR proche de 0.

Résultats et commentaires généraux. Ces figures illustrent que la méthode de *stability selection* est efficace. Notamment, les boxplots (figures 3.3, 3.4 et 3.5) montrent les distributions de la probabilité d'appartenir au modèle pour chacune des covariables. La différence de distributions entre les variables importantes X_1, X_2, X_3 et X_4 et les autres est claire. Les trois premières covariables sont quasiment tout le temps détectées avec une probabilité seuil $p_{thr} = 1$. La quatrième, dont le coefficient de régression β est le plus faible non nul, est moins bien détectée que les premières mais de façon bien meilleure que les covariables n'appartenant pas au modèle. Par exemple, pour le cas de la loi 3 (gaussiennes liées, figure 3.5), la covariable X_4 est détectée dans 75% des cas si $p_{thr} = 0.75$ pour $n = 200$ (respectivement $p_{thr} = 0.55$ pour $n = 100$).
Les courbes ROC (figure 3.6) permettent d'une part d'illustrer la différence entre les tailles

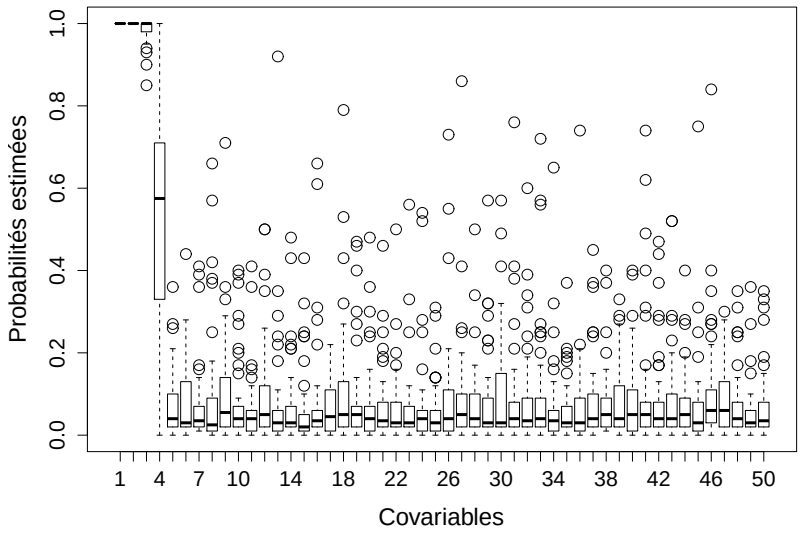


FIGURE 3.3 – Boxplots des probabilités estimées $\max_{\tau \in T} \hat{p}_k(\tau)$ pour chaque covariable. Les covariables sont indépendantes et gaussiennes (loi 1) et les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 50 répétitions constituées de $n = 100$ observations de $p = 50$ covariables avec la méthode de *stability selection*.

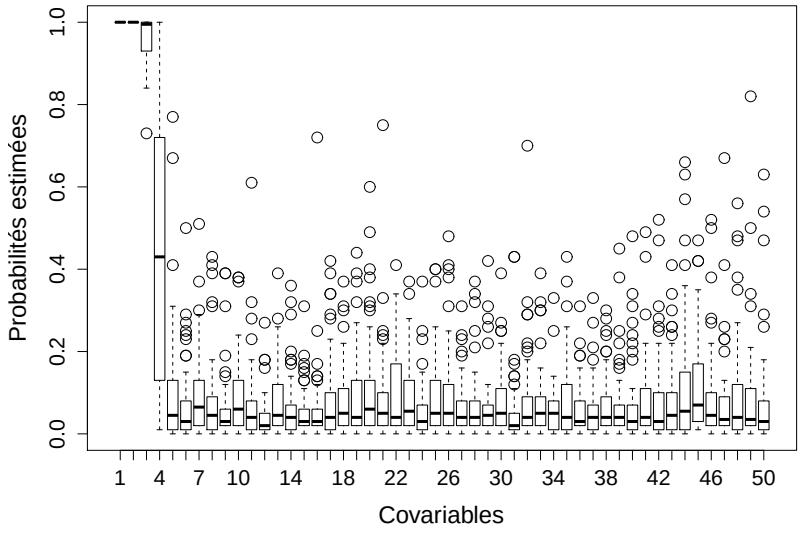
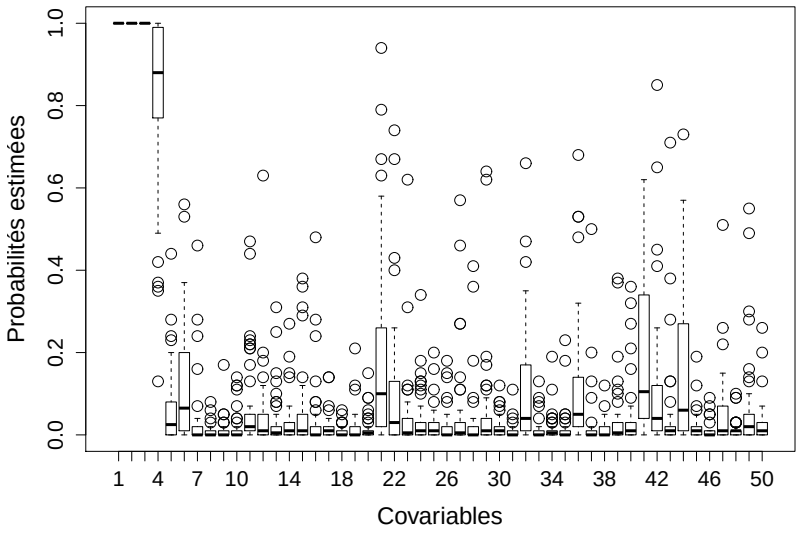
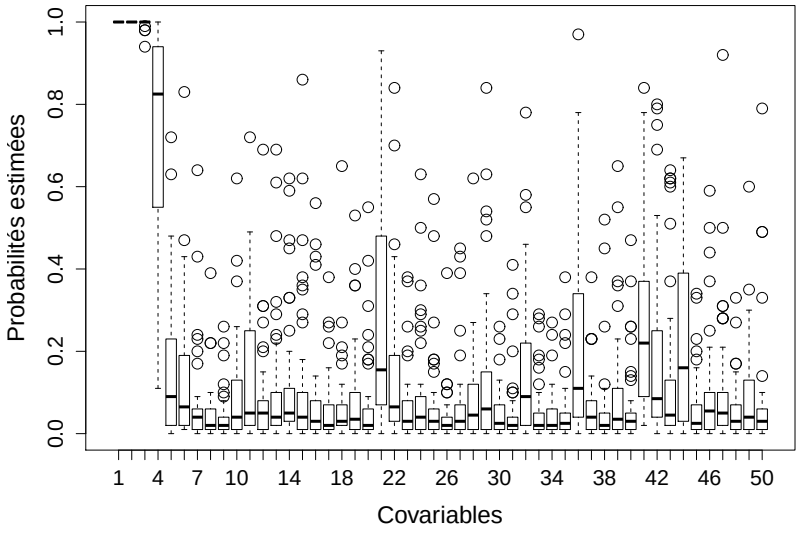


FIGURE 3.4 – Boxplots des probabilités estimées $\max_{\tau \in T} \hat{p}_k(\tau)$ pour chaque covariable. Les covariables sont indépendantes et distribuées selon le mélange de lois (loi 2). Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 50 répétitions constituées de $n = 100$ observations de $p = 50$ covariables avec la méthode de *stability selection*.



(a) $n = 200$



(b) $n = 100$

FIGURE 3.5 – Boxplots des probabilités estimées $\max_{\tau \in T} \hat{p}_k(\tau)$ pour chaque covariable. Les covariables sont des gaussiennes liées simulées selon la loi 3. Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 50 répétitions constituées de $n = 200$ et 100 observations de $p = 50$ covariables avec la méthode de *stability selection*.

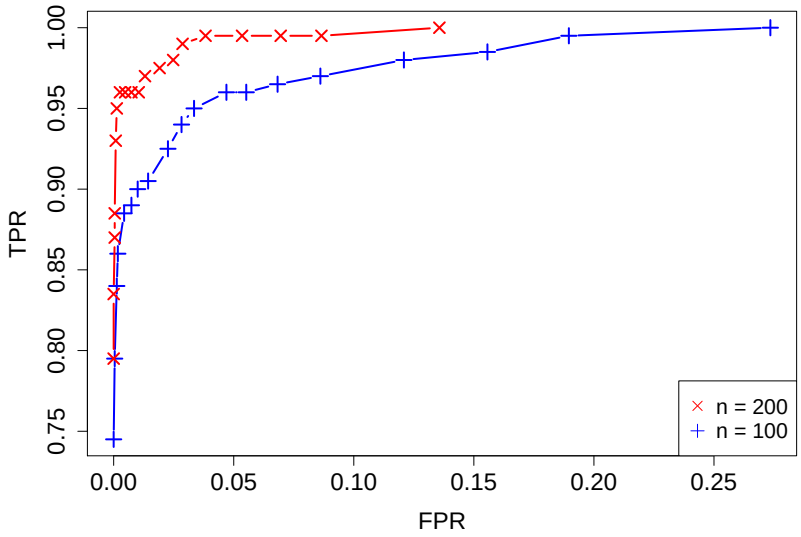


FIGURE 3.6 – Courbes ROC obtenues en seuillant pour $p_{thr} \in \{0.1, 0.15, 0.2, \ldots, 1\}$ avec la méthode de *stability selection*. Les covariables sont des gaussiennes liées simulées selon la loi 3. Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les TPR et FPR correspondent aux TPR et FPR moyens obtenus sur 50 répétitions de $n = 100$ et 200 observations de $p = 50$ covariables.

d'échantillons (en l'occurrence, $n = 200$ et $n = 100$ échantillons) et d'autre part, d'illustrer le phénomène de faible influence du seuil p_{thr} mentionné précédemment. Tout d'abord, les résultats sur la différence de taille d'échantillons sont assez prévisibles : la courbe ROC pour $n = 200$ est au-dessus de celle pour $n = 100$. En d'autres termes, plus la taille de l'échantillon est grande, meilleurs sont les résultats. Il est tout de même à noter que l'échelle pour le TPR commence à 0.75 et celle pour le FPR finit à 0.25. Ceci signifie que pour $p_{thr} = 1$ les TPR moyens sont respectivement d'environ 0.75 et 0.8 pour $n = 100$ et $n = 200$ et que pour $p_{thr} = 0.1$, les FPR moyens sont respectivement d'environ 0.27 et 0.14 pour $n = 100$ et $n = 200$. Dans les deux cas, les courbes ROC restent très satisfaisantes et montrent la sensibilité et spécificité de cette procédure. Ces deux courbes ROC permettent également d'illustrer qu'une grande étendue de valeurs de p_{thr} donne des résultats très similaires, justifiant l'intérêt de cette méthode qui permet d'éviter un choix du paramètre de pénalisation, beaucoup plus sensible à régler.
La différence entre les tailles d'échantillons est également illustrée par les boxplots de la figure 3.5. Sans surprise également, les résultats sont meilleurs pour l'échantillon le plus grand $n = 200$ où la covariable X_4 est mieux détectée tandis que les covariables n'appartenant pas au modèle sont moins détectées.

Comparaisons entre les différentes lois. Pour cela, il est judicieux de s'intéresser aux boxplots 3.3 et 3.4 dont les covariables sont dans les deux cas simulées de façon indépendantes. Ces deux boxplots ne présentent pas de grandes différences, la covariable X_4 semble un peu moins bien détectée avec le mélange de loi (figure 3.4). La méthode de détection par *stability selection* semble toutefois être assez robuste vis-à-vis de ce paramètre.

Comparaisons entre covariables indépendantes et covariables liées. Pour cette comparaison, on s'intéressera aux boxplots 3.3 et 3.5 (pour $n = 100$) dont les covariables sont des gaussiennes respectivement indépendantes et liées. On peut remarquer que la covariable X_4 est mieux détectée dans le cas lié, ceci est dû au fait que X_1 et X_4 sont dépendantes conditionnellement aux autres covariables. Le boxplot du cas lié (figure 3.5) montre des probabilités estimées globalement plus élevées concernant les covariables n'appartenant pas au modèle. Ceci est également en partie dû à la structure de dépendances conditionnelles sous-jacente, en effet, certaines de ces covariables sont liées à une des quatre premières covariables et sont par conséquent davantage détectées.

3.3.4 Knockoffs revisités

Dans ce qui suit, on va présenter des résultats de simulations concernant la méthode des knockoffs revisités. De la même façon que pour *stability selection*, on présentera dans un premier temps des résultats dans une configuration "favorable" à l'étude de l'efficacité de la méthode des knockoffs revisités, c'est-à-dire pour un relativement petit nombre de covariables et un assez grand nombre d'observations. Dans un second temps, on présentera des résultats pour un nombre de covariables plus conséquent avec moins d'observations pour montrer que la méthode présente encore des performances intéressantes dans cette configuration.

Configuration favorable

Cette première configuration est la même que celle exposée dans les résultats de simulations de la méthode de *stability selection* ; les données simulées utilisées sont d'ailleurs également les mêmes. On considère $n = 100$ et $n = 200$ observations de $p = 50$ covariables et $J = 3$ modalités pour la variable réponse Y, les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$ et les seuils $\boldsymbol{\alpha}$ choisis de façon à ce que la variable réponse Y prenne assez de valeurs dans chacune de ses trois modalités.

On représente les résultats par des boxplots (figures 3.7, 3.8 et 3.9) et taux de détection (figure 3.10). Ces figures sont toutes obtenues sur 100 répétitions de $n = 100$ ou $n = 200$ observations de $p = 50$ covariables. Les boxplots représentent les rangs "d'apparition" de chaque covariable, qui est une sorte de rang d'importance donné à l'issue de la première étape. Plus le rang est faible, plus la covariable est importante dans le modèle : la covariable dont le rang est 1 est la plus importante et est apparue la première dans le modèle etc. Avec ce système de rangs, des covariables peuvent avoir le même rang, ainsi le dernier rang attribué n'est pas forcément le p^{eme}. Ces rangs permettent de trier les covariables selon leur importance.
Le graphique des taux de détection (figure 3.10) présente le nombre de détection de chaque covariable sur 100 répétitions (donc le taux de détection en pourcentage) à l'issue de la deuxième étape. Les données utilisées sont les mêmes que celles utilisées pour générer la figure 3.9.

Résultats et commentaires généraux. Ces figures illustrent les bonnes performances de la méthode des knockoffs revisités. Comme attendu, les boxplots (figures 3.7, 3.8 et 3.9) indiquent que les covariables X_1, X_2, X_3 et X_4 entrent dans le modèle dans cet ordre. La différence entre les rangs de ces quatre covariables et les rangs du reste des covariables (qui n'appartiennent pas au modèle) est très nette.
Les taux de détection représentés dans la figure 3.10 corroborent également ce phénomène ; les trois premières covariables sont presque toujours détectées pour $n = 200$ comme pour $n = 100$

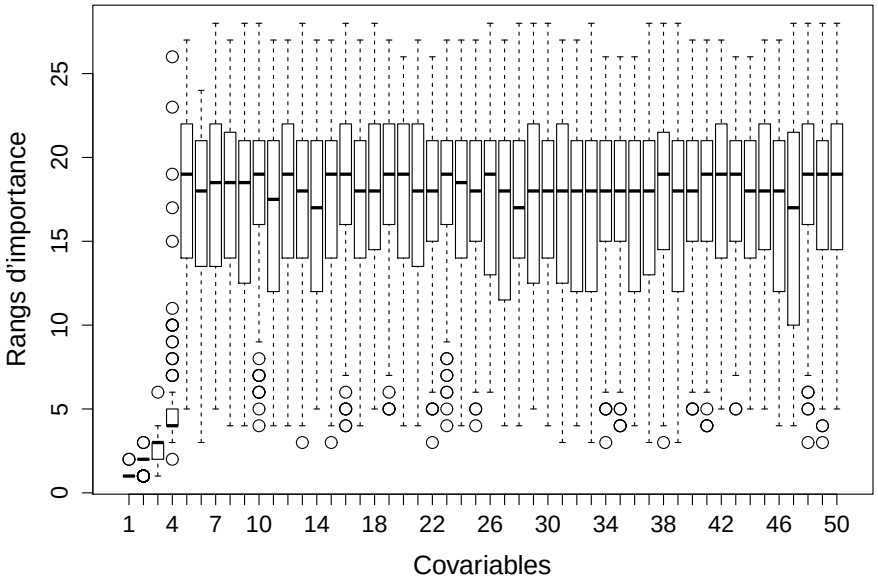


FIGURE 3.7 – Boxplots des rangs d'importance de chaque covariable obtenus avec la méthode des knockoffs revisités. Les covariables sont indépendantes et gaussiennes (loi 1) et les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 100 répétitions constituées de $n = 100$ observations de $p = 50$ covariables.

observations tandis que le taux de détection des covariables nulles est relativement faible (moins de 20%). La quatrième covariable est respectivement détectée dans 58% et 63% des cas pour $n = 100$ et $n = 200$ mais il s'agit de la covariable du modèle qui possède le plus petit coefficient de régression non nul. On peut toutefois observer que certaines covariables, comme X_{21}, X_{32}, X_{41} et X_{44}, sont un peu plus souvent détectées que les autres. Ceci est probablement dû à la structure de dépendance des covariables X. En particulier, ces covariables sont conditionnellement dépendantes à X_1, X_2, X_3 et X_4 qui appartiennent, elles, au modèle. On retrouve également ce phénomène sur la figure 3.9 (qui utilise les mêmes jeux de données), les covariables en question ayant un rang d'importance plus faible que les autres covariables nulles.
De la même façon que pour *stability selection*, on souhaite illustrer l'impact de différentes tailles d'échantillons (figures 3.10 et 3.9). Sur la figure 3.9, on peut observer que quand $n = 100$, les rangs des covariables nulles sont plus faibles sur certaines répétitions en comparaison à $n = 200$. Le rang de la covariable X_4 s'envole sur plusieurs répétitions pour $n = 100$, ce qui n'est pas le cas pour $n = 200$. La figure 3.10 illustre mieux cette différence : les covariables importantes sont un peu moins bien détectées pour $n = 100$ contrairement aux covariables nulles qui ont tendance à être un peu plus détectées pour $n = 100$. Les résultats sont donc un peu moins bons pour $n = 100$ que pour $n = 200$ comme attendu.

Comparaisons entre les différentes lois. De la même façon que précédemment, il est judicieux pour cela de comparer les figures 3.7 et 3.8 où les covariables sont simulées de façon indépendante dans les deux cas. Peu de différences sont à noter là encore. Sur ces boxplots, on remarque que les rangs médians des quatre premières covariables sont respectivement $1, 2, 3$ et 4.

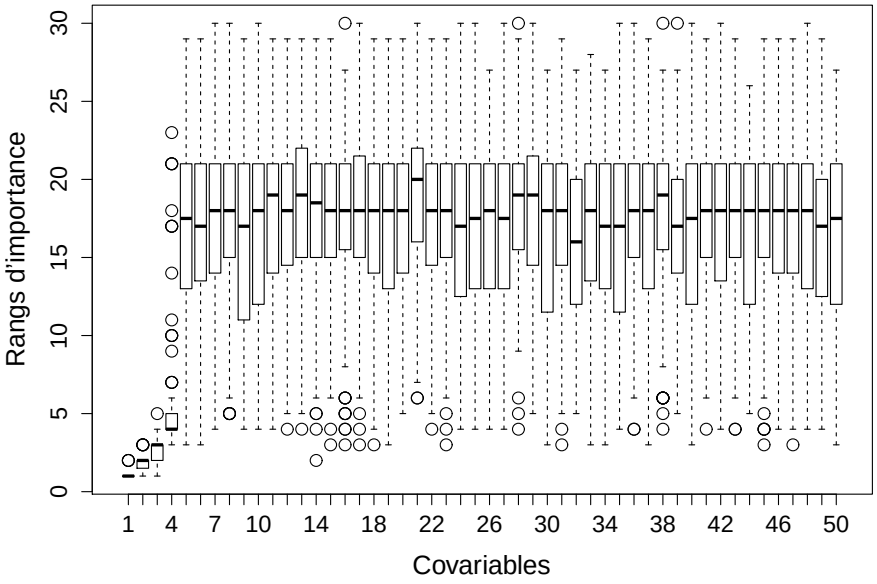
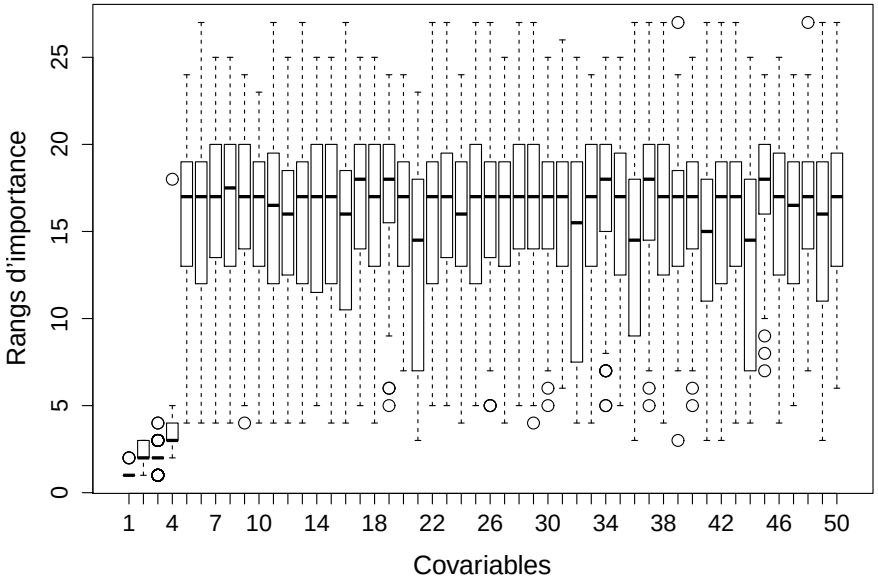


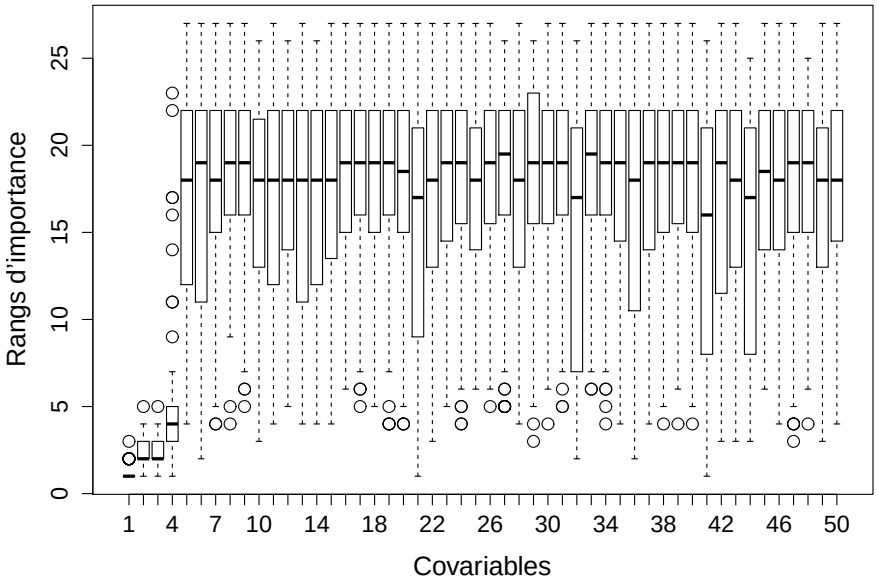
FIGURE 3.8 – Boxplots des rangs d'importance de chaque covariable obtenus avec la méthode des knockoffs revisités. Les covariables sont indépendantes et distribuées selon le mélange de lois (loi 2). Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 100 répétitions constituées de $n = 100$ observations de $p = 50$ covariables.

La covariable X_4 a un rang élevé sur plusieurs répétitions dans les deux cas et les rangs médians des covariables n'appartenant pas au modèle sont dans les deux cas entre 15 et 20.

Comparaisons entre covariables indépendantes et covariables liées. Pour déterminer l'impact de l'indépendance des covariables, on va comparer les figures 3.7 et 3.9 ($n = 100$) dont les covariables sont gaussiennes dans ces deux cas. Les données sont simulées de la même façon que pour *stability selection*. Ainsi on remarque le même phénomène, à savoir la covariable X_4 est un peu mieux détectée dans le cas lié (figure 3.9) car elle est conditionnellement dépendante à X_1 qui a un fort coefficient de régression. Le rang de X_4 dépasse 25 sur certaines répétitions dans le cas indépendant (figure 3.7), mais pas dans le cas lié. Pour les mêmes raisons, les rangs de certaines covariables n'appartenant pas au modèle ont tendance à être un peu plus faibles dans le cas lié (figure 3.9). C'est le cas par exemple des covariables X_{21}, X_{32}, X_{41} et X_{44} mentionnées précédemment qui sont conditionnellement dépendantes aux quatre premières covariables (qui appartiennent au modèle).



(a) $n = 200$



(b) $n = 100$

FIGURE 3.9 – Boxplots des rangs d'importance de chaque covariable obtenus avec la méthode des knockoffs revisités. Les covariables sont des gaussiennes liées simulées selon la loi 3. Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Les boxplots sont obtenus sur 100 répétitions constituées de $n = 100$ et 200 observations de $p = 50$ covariables.

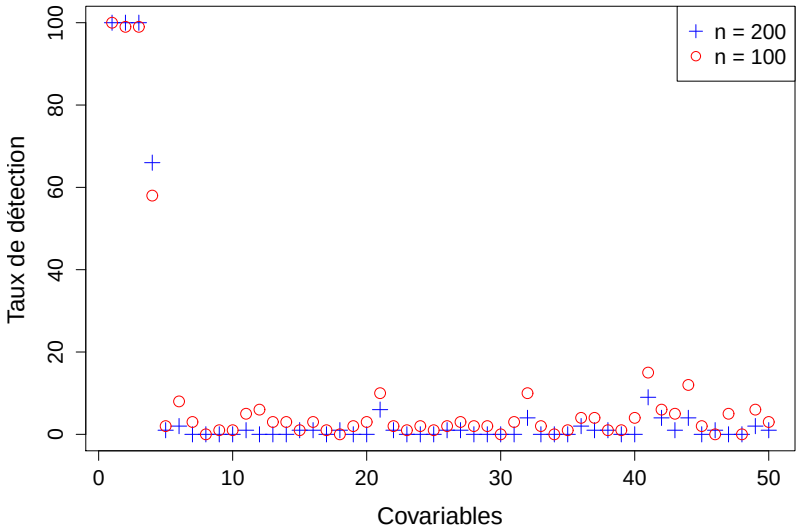


FIGURE 3.10 – Taux de détection de chaque covariable sur 100 répétitions après application de la méthode des knockoffs revisités. Les covariables sont des gaussiennes liées simulées selon la loi 3. Les coefficients de régression valent $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$.

Avec plus de covariables

À présent, on présente des résultats de simulations pour $n = 1000$ observations de $p = 2000$ covariables. La réponse Y possède ici aussi $J = 3$ modalités ordonnées, les coefficients de régressions valent $\beta_k = \begin{cases} 1, \text{ si } 1 \leq k \leq 20, \\ 0.75, \text{ si } 21 \leq k \leq 40, \\ 0.5, \text{ si } 41 \leq k \leq 60, \\ 0.25, \text{ si } 61 \leq k \leq 80, \\ 0, \text{ sinon,} \end{cases}$ et le vecteur $\boldsymbol{\alpha}$ des seuils est ici aussi choisi de sorte que la variable réponse Y prenne assez de valeurs dans chacune de ses trois modalités.

Pour davantage de lisibilité, on représente les résultats par les boxplots des taux de détection de chaque groupe de covariables (figures 3.11, 3.12, 3.13) simulées respectivement selon les lois 1, 2 et 3. Ces boxplots sont obtenus sur 100 répétitions de $n = 1000$ observations de $p = 2000$ covariables. Ces figures présentent les boxplots des taux de détection des cinq groupes de covariables (selon leur coefficient de régression β) après application de la méthode des knockoffs revisités.

Résultats et commentaires généraux. Les figures 3.11, 3.12 et 3.13 représentent les taux de détection pour les covariables simulées respectivement selon les lois 1, 2 et 3. Il apparaît clairement que ces taux de détection dépendent du coefficient de régression β : en effet, plus β est grand, plus les covariables associées sont détectées. Les taux de détection du premier groupe de covariables, pour lesquelles le coefficient de régression β vaut 1, sont très hauts : la moitié des covariables de ce groupe sont détectées 99 ou 100% sur les trois figures. Les taux de détection du second groupe sont un peu moins élevés que ceux du premier et ainsi de suite. La différence de taux de détection entre les trois premiers groupes de covariables, qui sont dans le modèle, est

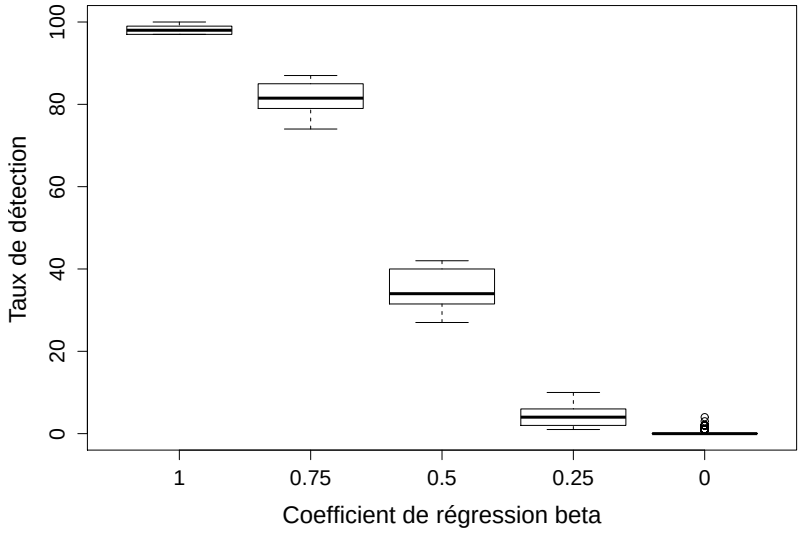


FIGURE 3.11 – Boxplots des taux de détection de chaque groupe de covariables en fonction de leur coefficient de régression $\beta \in \{1, 0.75, 0.5, 0.25, 0\}$. Les covariables sont indépendantes et gaussiennes (loi 1). Les boxplots sont obtenus sur 100 répétitions constituées de $n = 1000$ observations de $p = 2000$ covariables avec la méthode des knockoffs revisités.

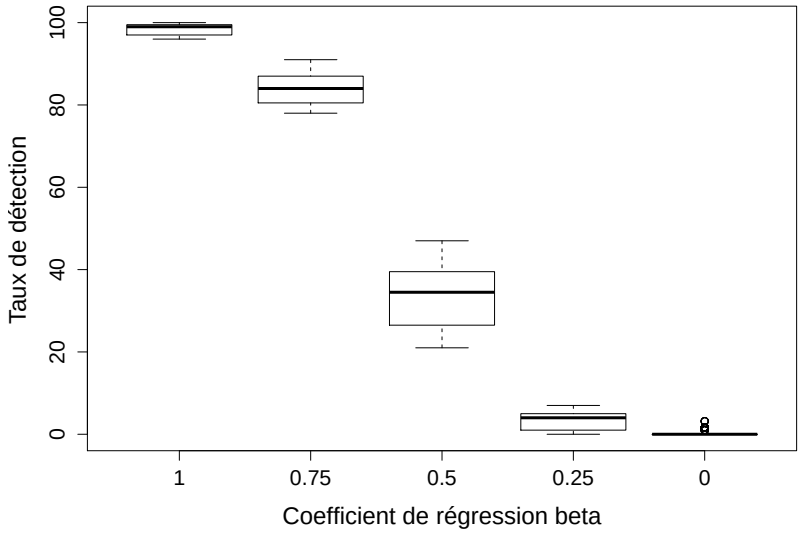


FIGURE 3.12 – Boxplots des taux de détection de chaque groupe de covariables en fonction de leur coefficient de régression $\beta \in \{1, 0.75, 0.5, 0.25, 0\}$. Les covariables sont indépendantes et distribuées selon le mélange de lois (loi 2). Les boxplots sont obtenus sur 100 répétitions constituées de $n = 1000$ observations de $p = 2000$ covariables avec la méthode des knockoffs revisités.

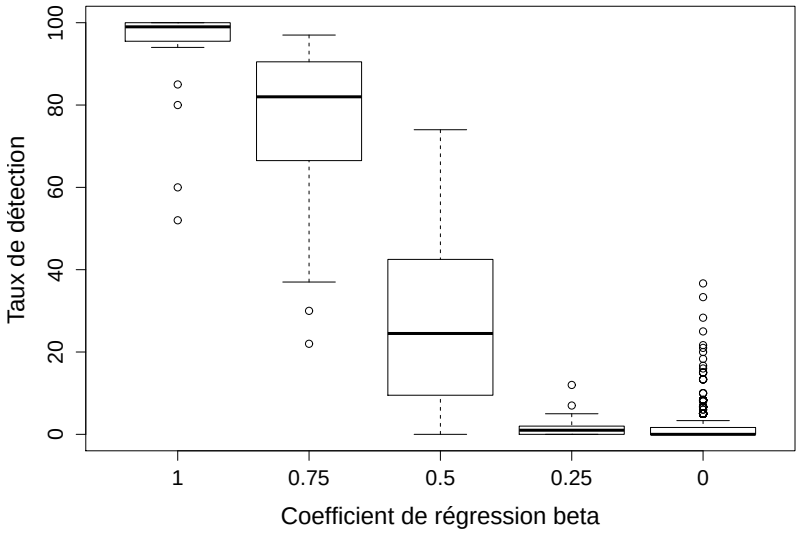


FIGURE 3.13 – Boxplots des taux de détection de chaque groupe de covariables en fonction de leur coefficient de régression $\beta \in \{1, 0.75, 0.5, 0.25, 0\}$. Les covariables sont des gaussiennes liées simulées selon la loi 3. Les boxplots sont obtenus sur 100 répétitions constituées de $n = 1000$ observations de $p = 2000$ covariables avec la méthode des knockoffs revisités.

claire en comparaison du dernier groupe de covariables, qui ne sont pas dans le modèle. Cependant, la différence entre le quatrième ($\beta = 0.25$) et le cinquième groupe ($\beta = 0$) n'est pas aussi évidente, en particulier sur la figure 3.13 des covariables gaussiennes liées. Ceci est probablement dû au faible coefficient de régression du quatrième groupe mais aussi à la structure sous-jacente de dépendances conditionnelles. En effet, les covariables de la figure 3.13 ne sont pas indépendantes et beaucoup de covariables qui n'appartiennent pas au modèle sont conditionnellement dépendantes à des covariables appartenant au modèle. Ce phénomène entraîne qu'elles sont alors davantage détectées.

Comparaisons entre covariables indépendantes et covariables liées. Par ailleurs, en comparant les figures 3.11 et 3.13 des covariables gaussiennes respectivement indépendantes et liées, on peut constater une plus grande variabilité des taux de détection. Sur la figure 3.11, les covariables du groupe 2 ont par exemple pour premier et troisième quartiles 78 et 82% contre 70 et 90% pour la figure 3.13. Ceci est probablement également dû à la structure de dépendances conditionnelles, qui modifie la détection selon "le voisinage" (les covariables auxquelles elles sont conditionnellement dépendantes). Pour les mêmes raisons, les covariables du dernier groupe sont moins détectées dans le cas indépendant (figure 3.11) que dans le cas lié (figure 3.13).

Comparaisons entre les différentes lois. Enfin, on peut remarquer qu'il n'y a pas de différences majeures entre les figures 3.11 et 3.12 (cas des covariables indépendantes de loi respectivement gaussienne et "mélange"). La distribution ne semble pas avoir de fort impact par rapport à la sélection de variables.

3.4 Conclusions

Dans ce chapitre, on a proposé une version Lasso de la régression logistique ordinale à logits cumulatifs qui s'appuie sur l'algorithme de Frank-Wolfe. Il se trouve qu'au cours de ce travail, une version Lasso de cette régression a été implémentée dans un package R, le package `ordinalNet`, par Michael Wurm (2017) [114]. Même si le temps d'exécution n'est pas encore optimal, ce package est assez complet et propose notamment d'autres types de pénalisations et d'autres régressions ordinales comme les *continuation-ratio models* ou les *adjacent-categories models* mentionnés dans l'introduction de cette partie. Les coefficients de régressions dans les simulations de ce chapitre ont malgré tout été estimés avec l'algorithme de Frank-Wolfe.
Pour la sélection de variables, on a développé une nouvelle méthode, la méthode des knockoffs revisités, qui est présentée en détails dans le chapitre suivant. Au vu des simulations, cette méthode semble efficace et pertinente sur ce modèle, indépendamment de différents paramètres comme le nombre d'observations et de covariables ou la distribution de celles-ci. En comparaison, on a également choisi d'essayer la méthode de validation croisée et la méthode de *stability selection*. Ces deux dernières méthodes sont relativement longues à exécuter, et la validation croisée a en plus tendance à sursélectionner. *Stability selection* se révèle en revanche plutôt efficace mais il faut être prudent au niveau des réglages du paramètre de pénalisation maximum τ_{max} qui peut facilement mener à de la sursélection.

La résolution du problème d'optimisation par l'algorithme de Frank-Wolfe est particulièrement efficace pour un objectif de sélection de variables. Cependant, la partie de l'optimisation sous contraintes liée à l'estimation des seuils $\boldsymbol{\alpha}$ est délicate ce qui rend l'estimation peu précise et donc inadaptée dans une version non pénalisée de cette régression. Cette estimation mériterait d'être améliorée, notamment dans un but de prédiction.
Par ailleurs, on a choisi de ne pas s'attarder sur l'étude de garanties théoriques concernant la méthode des knockoffs appliquée à ce modèle, notamment à cause de la complexité de ce dernier. Cependant, ceci serait intéressant et pourrait être l'objet de futurs travaux théoriques substantiels.

Chapitre 4

La méthode des knockoffs revisités pour la sélection de variables

Sommaire

4.1 Méthode des knockoffs revisités 67
4.1.1 Contexte 67
4.1.2 Principe et généralités 69
4.1.3 Choix du seuil 70
4.1.4 Package R `kosel` 72
4.2 Simulations 73
4.2.1 Paramètres de simulations 73
4.2.2 Efficacité et comparaisons - $p = 50$ 74
4.2.3 Efficacité et comparaisons - $p = 2000$ 78
4.2.4 Caractère aléatoire de la procédure 82
4.3 Conclusions 84

Dans ce chapitre, on considère le problème de sélection de variables dans des modèles de régression. On cherche à sélectionner les covariables explicatives liées à la variable réponse, c'est-à-dire à déterminer quelles covariables jouent un rôle important dans le modèle. On se restreint ici pour cela à des modèles de régression linéaire L_1-pénalisés. Pour gérer le choix du paramètre de pénalisation qui contrôle la sélection de variables, on a développé une nouvelle méthode basée sur l'idée des knockoffs de Barber et Candès (2015) [7] : la méthode des knockoffs revisités. Cette méthode, introduite brièvement au chapitre 3, présente de bonnes performances sur le modèle de régression à logits cumulatifs. Il s'avère que cette méthode est plus générale et convient en fait à un spectre plus large de régressions pour des types de variables réponses variés. De plus, elle fonctionne également quand le nombre d'observations est plus petit que le nombre de covariables et permet de trier les covariables selon leur ordre d'importance.

4.1 Méthode des knockoffs revisités

4.1.1 Contexte

On suppose qu'on a p covariables explicatives à valeurs réelles $\overrightarrow{X} := (X_1, X_2, \ldots, X_p)$ et une variable réponse Y liée au vecteur $\overrightarrow{X}$ de covariables par m équations du type :

$$f_k(\mu_k(Y|X)) = \alpha_k + \beta_1 X_1 + \ldots + \beta_p X_p, \ k = 1, \ldots, m, \tag{4.1}$$

où f_k est une fonction déterministe connue, $\mu_k(Y|X)$ un paramètre de la loi conditionnelle de Y sachant X et $\alpha_k, \beta_1, \ldots, \beta_p$ des coefficients réels. Le vecteur $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)$ des coefficients de régression de dépend pas de k.

Ce cadre est assez général et englobe plusieurs modèles de régression (linéaire) tels que les modèles linéaires généralisés [104, 67, 2, 107], introduits à la sous-section 2.1.2 du chapitre 2, et dont les régressions linéaire, logistique, multinomiale et de Poisson font notamment partie. Ce cadre englobe également les modèles de régression ordinale logistique [1] : *cumulative logit models* [89, 112, 4], *adjacent-categories logit models*, *continuation-ratio logit models*, mais aussi les *cumulative link models* [1]. En effet, pour les modèles linéaires généralisés, $m = 1$, $\mu_1(Y|X) = \mathbb{E}(Y|X)$ et f_1 est la fonction de lien du modèle correspondant (identité pour la régression linéaire, log pour la régression de Poisson, logit pour la régression logistique...). Pour les modèles de régression ordinale logistique, $f_k = \text{logit}$ et $\mu_k(Y|X) = \mathbb{P}(Y \leq k|X)$ (*cumulative logit*), $\mu_k(Y|X) = \mathbb{P}(Y = k|Y = k \text{ ou } Y = k+1, X)$ (*adjacent-categories logit*) $\mu_k(Y|X) = \mathbb{P}(Y > k|Y \geq k, X)$ (*continuation-ratio logit*). La régression à logits cumulatifs (*cumulative logit model*) correspond au modèle à odds proportionnels présenté au chapitre 3. Pour les *cumulative link models*, la fonction f_k de lien change (souvent la fonction probit ou la fonction log-log). Ces derniers modèles ne permettent que les effets identiques des covariables, c'est-à-dire que les coefficients de régression β_j ne dépendent pas de la modalité k de la variable réponse Y. Ce cadre comprend donc des modèles pour différents types de variable réponse, notamment binaire, continue, catégorielle et ordinale.

Dans ce contexte, les covariables $X_l,\ l = 1, \ldots, p$ doivent être liées à la variable réponse Y par une expression linéaire afin que la dépendance conditionnelle entre Y et X_l sachant $X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_p$ puisse être mesurée par le coefficient de régression β_l. Plus précisément, $\beta_l = 0$ signifie que X_l et Y sont indépendantes conditionnellement aux autres covariables $X_k,\ k = 1, \ldots, l-1, l+1, \ldots p$. On s'intéresse donc à la nullité des coefficients de régression β_l pour sélectionner les covariables importantes. De plus, on fait une hypothèse de parcimonie, c'est-à-dire qu'on suppose qu'un relativement petit nombre de covariables jouent un rôle important dans le modèle. Ceci implique que seulement peu de covariables sont importantes et donc, que seulement un faible nombre de coefficients de régression β_l sont non nuls. Cette hypothèse de parcimonie est commode pour les scientifiques afin de restreindre leurs études à un plus petit sous-ensemble de covariables, notamment dans un contexte de grande dimension. Au lieu de vérifier la nullité de chacun des coefficients de régression β_l à l'aide de tests statistiques, on ajoute une pénalisation en norme L_1 sur le vecteur $\boldsymbol{\beta}$ des coefficients de régression dans l'estimation des coefficients du modèle. Les coefficients sont ordinairement estimés comme la solution du problème d'optimisation :

$$\underset{(\boldsymbol{\alpha},\boldsymbol{\beta})}{\text{argmax}}\ L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}), \tag{4.2}$$

où $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X})$ est une fonction des coefficients du modèle (souvent la log-vraisemblance), qui dépend des observations $\mathbf{Y}$ et $\mathbf{X}$ de la variable réponse Y et du vecteur de covariables $\overrightarrow{X}$ respectivement. Au lieu d'estimer les coefficients en résolvant (4.2), on ajoute une pénalisation Lasso sur le vecteur $\boldsymbol{\beta}$ des coefficients de régression, ce qui revient à résoudre le problème d'optimisation suivant :

$$\underset{(\boldsymbol{\alpha},\boldsymbol{\beta})}{\text{argmax}}\ \{L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}) - \lambda||\boldsymbol{\beta}||_1\}, \tag{4.3}$$

où $\lambda > 0$ est le paramètre de pénalisation.

Les méthodes de pénalisation requièrent ensuite le choix du paramètre de pénalisation, parfois aussi appelé paramètre de régularisation (*regularisation parameter*) ou paramètre de réglage (*tuning parameter*). On a donc besoin ici de régler ce paramètre de pénalisation $\lambda > 0$ qui apparaît dans le problème d'optimisation (4.3). Ce paramètre contrôle le nombre de covariables sélectionnées : plus λ est grand, moins il y a de covariables sélectionnées. Au contraire, des valeurs de λ proches de 0 conduisent au modèle plein, c'est-à-dire au modèle avec toutes les covariables.

Remarque : Dans le chapitre 3, on travaille avec un paramètre de pénalisation τ différent, impliqué dans un problème d'optimisation équivalent à (4.3). L'équivalence de ces problèmes d'optimisation a également été évoquée dans la sous-section 2.1.2 du chapitre 2. Bien qu'il y ait une correspondance bijective entre ces deux paramètres de pénalisation $\tau > 0$ et $\lambda > 0$, il faut avoir en tête que cette correspondance est inversement proportionnelle : des valeurs proches de 0 de τ correspondent à des valeurs élevées de λ et vice versa.

Notre but est de sélectionner seulement les covariables importantes et donc, d'éviter les faux positifs (les covariables sélectionnées à tort). Pour répondre à ce problème, on propose une nouvelle méthode, inspirée de la méthode des knockoffs de Barber et Candès (2015) [7] dans le cadre de la régression linéaire gaussienne. En fait, cette méthode ne donne pas directement une valeur du paramètre de pénalisation λ mais elle trie les covariables de la plus susceptible d'être importante dans le modèle à la moins susceptible de l'être. En outre, elle convient à n'importe quelle régression du type présenté en (4.1) y compris quand le nombre n d'observations est plus petit que le nombre p de covariables. Dans le modèle linéaire gaussien, il est toutefois plus pertinent d'utiliser la procédure décrite dans [7] à cause de leurs garanties théoriques. Même si leur procédure était initiallement valable pour $n > p$, ils l'ont par la suite étendue au cas $n \leq p$ grâce à une étape préliminaire de filtrage [8]. Dans ce qui suit, on présente le principe de notre méthode des knockoffs revisités.

4.1.2 Principe et généralités

On note $\mathbf{X}$ la matrice $n \times p$ des n observations du p-vecteur $\overrightarrow{X} = (X_1, \ldots, X_p)$ de covariables, appelée la matrice design. Le principe, développé par [7], est d'utiliser une matrice $\tilde{\mathbf{X}}$ de copies (knockoffs) des covariables X_i dont la structure de covariance est similaire à celle de $\mathbf{X}$ mais indépendante de $\mathbf{Y}$. On fait progressivement décroître le paramètre de pénalisation λ de $+\infty$ à 0 et on va déterminer la plus grande valeur de λ pour laquelle chaque covariable et chaque copie entre dans le modèle, c'est-à-dire la plus grande valeur de λ pour laquelle le coefficient de régression estimé associé est non-nul. Le but est de déterminer si une covariable X_i appartient au modèle en étudiant si elle entre dans le modèle avant sa copie $\tilde{X}_i$, c'est-à-dire si X_i entre dans le modèle pour une valeur du paramètre λ supérieure à celle de sa copie. En effet, comme la matrice $\tilde{\mathbf{X}}$ des copies est construite de manière à être indépendante des observations $\mathbf{Y}$ de la variable réponse, si une covariable entre dans le modèle après sa copie, on peut légitimement suspecter que cette covariable n'appartient pas au modèle.

La principale différence de notre méthode avec celle de Barber et Candès [7] réside dans la construction de la matrice des knockoffs $\tilde{\mathbf{X}}$. Dans leur papier, ils proposent une construction sophistiquée de cette matrice à l'aide d'outils d'algèbre linéaire. Cette construction permet de contrôler le *false discovery rate* (FDR), la proportion moyenne de faux positifs parmi les positifs, dans le modèle linéaire gaussien quand il y a au moins autant d'observations que de covariables

($n \geq p$). Cette différence dans la construction des copies rend notre méthode utilisable dans le cas $n < p$ et pour un plus grand nombre de modèles de régression. Cependant, les garanties théoriques concernant le contrôle du *false discovery rate* ne sont plus valables.
On construit notre matrice $\tilde{\mathbf{X}}$ de copies des covariables en permutant aléatoirement les n lignes de la matrice design $\mathbf{X}$. De cette façon, les corrélations entre les copies demeurent identiques à celles entre les variables d'origine mais les copies ne sont plus liées à la variable réponse $\mathbf{Y}$. Cette construction de la matrice de copies rend la procédure aléatoire.

Ensuite, de la même façon que [7], on effectue la régression de $\mathbf{Y}$ sur la matrice augmentée $[\mathbf{X}, \tilde{\mathbf{X}}]$ de taille $n \times 2p$, qui correspond à la concaténation des matrices $\mathbf{X}$ (de taille $n \times p$) et $\tilde{\mathbf{X}}$ (de taille $n \times p$) par les colonnes. On note $\hat{\boldsymbol{\beta}}(\tau)$ les coefficients estimés de la régression λ-pénalisée de la variable réponse $\mathbf{Y}$ sur la matrice augmentée $[\mathbf{X}, \tilde{\mathbf{X}}]$:

$$\left(\hat{\boldsymbol{\alpha}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda)\right) := \underset{(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \left\{ L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]) - \lambda ||\boldsymbol{\beta}||_1 \right\}.$$

Pour chaque variable de la matrice augmentée, c'est-à-dire pour chaque covariable et copie correspondante, on considère la statistique $T_i := \sup \{\lambda > 0,\ \hat{\beta}_i(\lambda) \neq 0\},\ i \in \{1, \ldots, p, p+1, \ldots, 2p\}$. La statistique T_i correspond à la plus grande valeur de λ pour laquelle la covariable X_i (respectivement la copie $\tilde{X}_{i-p}$ de X_i) si $i \in \{1, \ldots, p\}$ (respectivement si $i \in \{p+1, \ldots, 2p\}$) entre dans le modèle estimé pour la première fois. On espère alors à ce moment que T_i est grand pour les variables importantes (celles appartenant au modèle), c'est-à-dire pour les variables X_i, $i \in \{1, \ldots, p\}$ telles que $\beta_i \neq 0$, et petit pour les copies $\tilde{X}_{i-p},\ i \in \{p+1, \ldots, 2p\}$ ou pour les covariables nulles, i.e. les covariables $X_i,\ i \in \{1, \ldots, p\}$ telles que $\beta_i = 0$ (qui n'appartiennent donc pas au modèle). Ceci fournit un $2p$-vecteur $(T_1, ..., T_p, \tilde{T}_1, ..., \tilde{T}_p)$ où $\tilde{T}_i$ désigne T_{i+p}. On considère alors, pour tout $i \in \{1, ..., p\}$, la statistique $W_i := \max(T_i, \tilde{T}_i) \times \begin{cases} (+1) & \text{si } T_i > \tilde{T}_i \\ (-1) & \text{si } T_i \leq \tilde{T}_i \end{cases}$.

Cette statistique permet de mesurer si une variable entre dans le modèle avant ou après sa copie : une valeur négative de W_i indique en effet que la covariable X_i est entrée dans le modèle après sa copie et on fait le choix de l'éliminer. À l'inverse, une valeur positive de W_i signifie que la covariable X_i est entrée dans le modèle avant sa copie et est donc plus susceptible d'appartenir au modèle. Toutefois, les covariables X_i dont la statistique W_i est positive n'appartiennent pas nécessairement au modèle : on espère que W_i est grand pour la plupart des covariables importantes et petit pour les covariables nulles. On est ainsi intéressé par les plus grandes valeurs positives du p-vecteur de statistiques W qui indiquent de plus que la variable est entrée "tôt" dans le modèle, en d'autres termes pour une grande valeur du paramètre de pénalisation λ. Ces statistiques W_i permettent de trier les covariables en fonction de leur importance dans le modèle : plus W_i est grand, plus la covariable associée X_i est susceptible d'être importante dans le modèle. Ce constat nous pousse à définir un seuil $s > 0$ pour W_i au-dessus duquel on gardera les covariables correspondantes dans le modèle estimé. Finalement, le modèle estimé $\hat{S}$ est donné par :

$$\hat{S} := \{X_i : W_i \geq s\}.$$

4.1.3 Choix du seuil

La deuxième différence majeure avec Barber et Candès [7] concerne le choix du seuil s. Ils fournissent un seuil dépendant des données qui montre des résultats assez attractifs par rapport au *false discovery rate* pour la régression linéaire gaussienne. Malheureusement, ces résultats ne

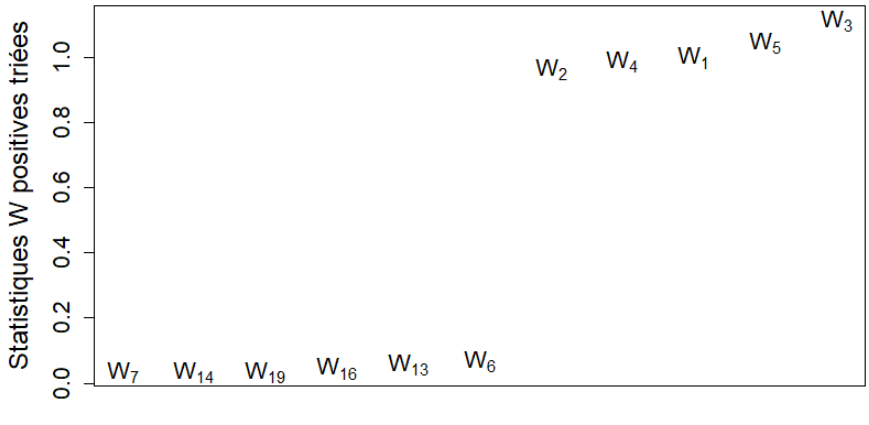


FIGURE 4.1 – Exemple de statistiques positives W_i triées dans l'ordre croissant. Régression linéaire gaussienne avec $n = 500$ observations de $p = 20$ covariables. Seules les covariables X_1, X_2, X_3, X_4 et X_5 appartiennent au modèle (dans ce cas, les coefficients de régression valent $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$).

sont plus valables en général, en-dehors du modèle linéaire gaussien. Pour notre méthode, on fait l'hypothèse qu'il y a une rupture entre les distributions des W_i correspondant aux covariables X_i dans le modèle et les autres (voir figure 4.1 à titre d'illustration). La figure 4.1 illustre que les distributions des W_i semblent être différente selon si la covariable associée est importante ou nulle. Pour générer la figure 4.1, on a simulé un jeu de données sous la modèle de régression linéaire avec $p = 20$ covariables indépendantes gaussiennes. Seulement les cinq premières sont liées à Y et donc, appartiennent au modèle :

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

où $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$ et $\epsilon \sim \mathcal{N}(0, 1)$. Dans notre procédure des knockoffs sur ce jeu de données, seules les statistiques $W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_{13}, W_{14}, W_{16}, W_{19}$ associées aux covariables $X_1, X_2, \ldots, X_{14}, X_{16}$ et X_{19} ont une valeur positive. On peut par exemple lire que la covariable X_1 est entrée dans le modèle pour $\lambda = 1.002$ (et donc, $T_1 = 1.002$) et déduire qu'elle est entrée dans le modèle avant sa copie $\tilde{X}_1$. Cela signifie que sa copie $\tilde{X}_1$ est entrée dans le modèle pour $\lambda < 1.002$ et entraîne que $\tilde{T}_1 < 1.002$. W_3 a la plus grande valeur parmi les statistiques $W_i,\ i = 1, \ldots, 20$, ce qui implique que X_3 est la covariable la plus susceptible d'appartenir au modèle. On peut clairement observer une rupture entre les valeurs des statistiques des cinq premières covariables et les autres.

En conséquence, on présente deux façons automatisées de choisir le seuil s à l'aide de deux méthodes de détection de ruptures : la méthode proposée par Auger et Lawrence (1989) [5] (et appliquée et améliorée par Picard *et al.* (2004) [78] et (2007) [79]) et la méthode CUSUM de détection de ruptures sur la moyenne. Notons $W_{(i)},\ i = 1, \ldots, w$ les w statistiques positives $W_i,\ i = 1, \ldots, w$ triées dans l'ordre croissant, telles que $0 < W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(w)}$ et $e_i = W_{(i+1)} - W_{(i)},\ i = 1, \ldots w - 1$ les $w - 1$ écarts entre ces statistiques triées. On remarque que w, le nombre de statistiques positives W_i, est aléatoire (sur la figure 4.1, $w = 11$). On propose deux seuils définis comme suit :

— le minimum des deux seuils obtenus en appliquant ces deux méthodes de détection de ruptures directement sur les statistiques $W_{(i)},\ i = 1, \ldots, w$ triées dans l'ordre croissant,

— le minimum des deux seuils obtenus en appliquant ces deux méthodes de détection de ruptures sur les écarts $e_i,\ i = 1, \ldots, w-1$.

Pour simplifier la lecture, on appellera le premier seuil "seuil W" et le second "seuil gaps".

4.1.4 Package R `kosel`

Nos procédures ont été implémentées dans un package R, appelé `kosel` [53] (pour *knockoffs selection*), disponible sur le CRAN. Notre package comprend trois fonctions : `ko.glm`, `ko.ordinal` et `ko.sel`.

Les deux premières fonctions construisent la matrice des knockoffs (copies des covariables) et retournent le p-vecteur de statistiques W pour les modèles de régressions L_1-penalisés implémentés dans les fonctions R `glmnet` et `ordinalNet` issues des packages du même nom. `glmnet` implémente les modèles linéaires généralisés et `ordinalNet` les modèles de régressions ordinales tels que *cumulative link*, *adjacent*, *continuation-ratio* ou encore *stopping ratio models*. Par défaut, une seed est utilisée pour que la matrice des knockoffs reste la même (et donc, le vecteur de statistiques W) lorsqu'on effectue différents runs sur un même jeu de données. Mais ceci peut être modifié avec l'option `random = TRUE` pour exploiter le caractère aléatoire de la procédure (voir sous-section 4.2.4 pour davantage de détails).

La troisième fonction, `ko.sel`, gère le choix du seuil. Elle prend en paramètre le vecteur de statistiques W retourné par une des deux précédentes fonctions et retourne le p-vecteur binaire correspondant au modèle estimé et le seuil s. Trois choix sont proposés : `method = 'stats'` et `method = 'gaps'` correspondent respectivement aux seuils W et gaps alors que l'option `method = 'manual'` permet à l'utilisateur de choisir son propre seuil. L'option `print = TRUE` affiche les statistiques positives W_i triées dans l'ordre croissant comme sur la figure 4.1. Pour `method = 'manual'`, elles sont automatiquement affichées pour que l'utilisateur puisse faire son choix. Pour les deux autres seuils "seuil W" (`method = 'stats'`) et "seuil gaps" (`method = 'gaps'`), l'option `print = TRUE` affiche également une ligne horizontale correspondant au seuil (voir figure 4.2).

Avec l'exemple des données de la figure 4.1, simulées sous le modèle de régression linéaire gaussienne, on utilise le package de la façon suivante : on calcule dans un premier temps les statistiques W à l'aide de la fonction `ko.glm`, puis on applique la fonction `ko.sel` à ces statistiques pour obtenir le seuil et le modèle estimé. Le code est le suivant :

```
> W = ko.glm(X, Y, family = "gaussian")
> ko.sel(W, print = TRUE, method = 'stats')
$estimation
     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
[1,] 1 1 1 1 1 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0

$threshold
[1] 0.9661426
```

La fonction `ko.sel` retourne le seuil W (car `method = 'stats'`), qui vaut dans ce cas `$threshold = 0.9661426` et le vecteur `$estimation` correspondant au modèle estimé c'est-à-dire aux covariables X_i dont la statistique W_i est supérieure ou égale au seuil `$threshold` et qui correspond dans ce cas aux cinq premières covariables. Comme `print = TRUE`, la fonction `ko.sel` retourne également le graphique de la figure 4.2 qui correspond aux statistiques positives W_i triées dans l'ordre croissant. La ligne horizontale rouge correspond au seuil `$threshold`. Les covariables dont la statistique W_i est au-dessus (au sens large) de cette ligne sont les covariables du modèle estimé.

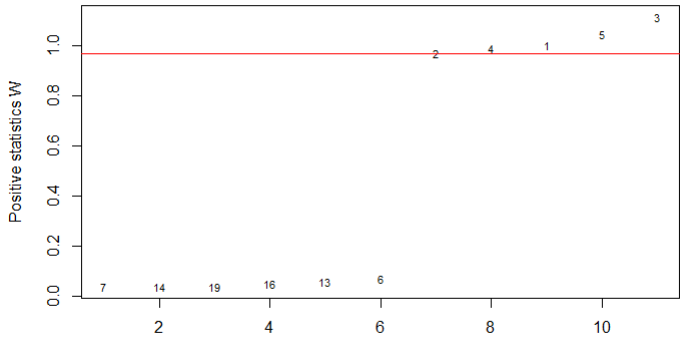


FIGURE 4.2 – Exemple d'affichage de la fonction `ko.sel`. Les données sont les mêmes que celles de la figure 4.1. Régression linéaire gaussienne avec $n = 500$ observations de $p = 20$ covariables. Seules les covariables X_1, X_2, X_3, X_4 et X_5 appartiennent au modèle (les coefficients de régression valent $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$).

4.2 Simulations

4.2.1 Paramètres de simulations

On expose à présent quelques résultats expérimentaux pour étudier l'efficacité de nos procédures. Pour cela, nous avons effectué différentes simulations avec des régressions variées : la régression linéaire gaussienne, la régression logistique et la régression ordinale à logits cumulatifs (et odds proportionnels). Les covariables $\overrightarrow{X}$ sont des gaussiennes centrées réduites ($\mathbb{E}(X_k) = 0$ et $\text{var}(X_k) = 1$ pour tout $k = 1, \ldots, p$) telles que X_i et X_j sont dépendantes conditionnellement aux autres covariables X_k, $k \in \{1, \ldots, p\} \setminus \{i, j\}$ avec probabilité 0.2. La matrice design $\mathbf{X}$ des covariables a été simulée avec la fonction R `huge.generator` du package `huge`, pour une structure de graphe `'random'`. On a ensuite simulé les observations de la variable réponse Y comme :

$$\begin{aligned} Y &= \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, && \text{(régression linéaire)} \\ \text{logit}(\mathbb{P}(Y = 1|X)) &= \alpha_1 + \beta_1 X_1 + \ldots + \beta_p X_p, && \text{(régression logistique)} \\ \text{ou logit}(\mathbb{P}(Y \leq k|X)) &= \alpha_k + \beta_1 X_1 + \ldots + \beta_p X_p, k = 1, 2, && \text{(régression à logits cumulatifs)} \end{aligned}$$

où $\epsilon \sim \mathcal{N}(0, 1)$ est du bruit gaussien, le vecteur $\boldsymbol{\beta}$ des coefficients de régression est parcimonieux et donné ci-dessous et les seuils α_1 et α_2 sont choisis de sorte que la variable réponse $\mathbf{Y}$ ait un nombre raisonnable d'observations pour chacune de ses modalités ($\{0, 1\}$ pour la régression logistique et $\{0, 1, 2\}$ pour la régression à logits cumulatifs). Les paramètres de ces régressions ont été respectivement estimés avec les fonctions R `glmnet` et `ordinalNet` des packages éponymes.

On présente les taux de détection de chacune des covariables sur $B = 100$ répétitions i.i.d. pour différentes configurations. Le taux de détection de la covariable X_l est le nombre de fois parmi les 100 répétitions où le modèle estimé contenait X_l. On a d'abord simulé $n = 200$ observations de $p = 50$ covariables pour des raisons pédagogiques et soucis de lisibilité de graphiques. Puis, on a simulé $n = 1000$ observations de $p = 2000$ covariables pour illustrer l'efficacité de la

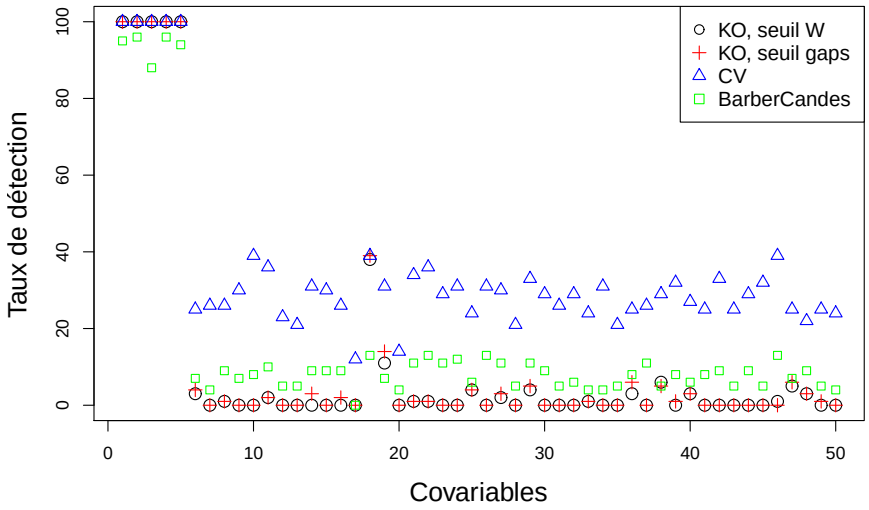
procédure dans un cadre de plus grande dimension. Pour $p = 2000$ covariables, les résultats sont présentés sous forme de boxplots de taux de détection en fonction du coefficient de régression β afin d'améliorer la lisibilité de ces graphiques.

On compare nos résultats à ceux obtenus avec la validation croisée (souvent abrégée en CV pour *cross validation*). La validation croisée a été effectuée avec les fonctions R `cv.glmnet` pour les régressions linéaire et logistique et `ordinalNetTune` avec la méthode `'logLik'` pour la régression à logits cumulatifs. Pour $p = 50$ et dans le cas de la régression linéaire gaussienne, on compare aussi nos résultats à ceux obtenus avec la procédure de Barber et Candès. Leur procédure est implémentée dans la fonction R `knockoff.filter` du package `knockoff`. On n'effectue pas cette comparaison pour $p = 2000$ en raison du trop faible nombre d'observations (qui rend leur procédure inapplicable).

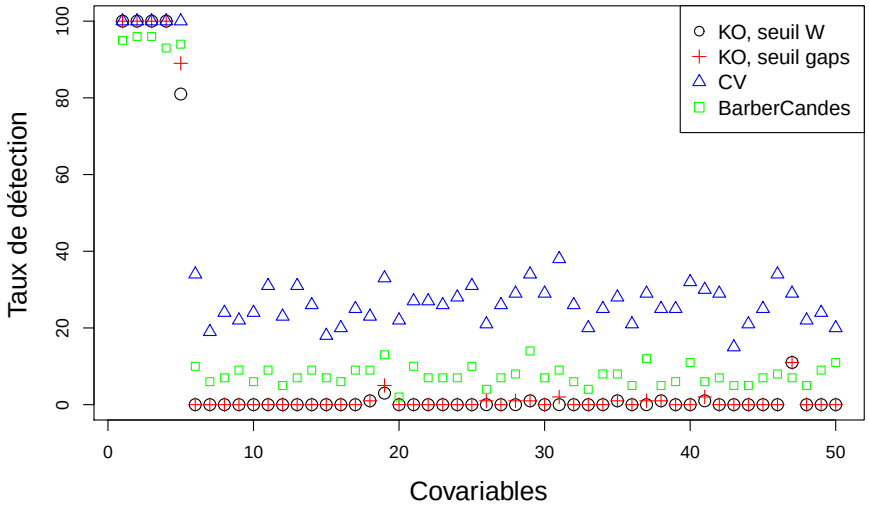
4.2.2 Efficacité et comparaisons - $p = 50$

Dans un premier temps, on présente les résultats pour $n = 200$ observations de $p = 50$ covariables et pour deux valeurs du vecteur des coefficients de régression : $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ et $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$. La matrice des covariables $\mathbf{X}$ reste la même pour chacune des régressions mais diffère selon le vecteur $\boldsymbol{\beta}$ des coefficients de régression. En d'autres termes, pour une valeur fixée de $\boldsymbol{\beta}$, on a gardé la même matrice design $\mathbf{X}$ pour simuler chacune des trois régressions. La variable réponse $\mathbf{Y}$ est ensuite simulée selon le modèle de régression à partir de cette matrice design et est donc différente selon les régressions. La matrice des knockoffs est également différente.

Résultats et commentaires. Les figures 4.3, 4.4 et 4.5 montrent les taux de détection pour la validation croisée et pour la méthode des knockoffs revisités avec le seuil W et avec le seuil gaps. Ces taux de détection sont présentés sur les figures 4.3, 4.4 et 4.5 pour les régressions linéaire, logistique et à logits cumulatifs respectivement. On peut d'abord remarquer que notre procédure est efficace pour chacun de ces trois modèles de régression : la différence de taux de détection entre les cinq premières covariables et le reste des covariables est vraiment nette, indépendamment des coefficients de régression (que ce soit pour $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ ou pour $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$) ou du choix du seuil. Pour la régression linéaire (figure 4.3), ces deux seuils donnent des résultats similaires alors que pour les régressions logistique (figure 4.4) et à logits cumulatifs (figure 4.5), le seuil gaps a tendance à donner des taux de détection légèrement inférieurs à ceux obtenus avec le seuil W, pour les covariables importantes comme pour les covariables nulles.
En comparaison, les taux de détection obtenus avec la validation croisée sont considérablement supérieurs : bien que les cinq premières covariables, particulièrement X_4 et X_5, sont mieux détectées, les covariables nulles le sont également davantage qu'avec nos procédures. Par exemple, pour la régression logistique (figure 4.4), les covariables nulles sont presque toujours détectées moins de 20% avec nos procédures alors qu'elles sont détectées entre 20% et 40% avec la validation croisée. En pratique, l'utilisation de la validation croisée risque de donner plus de faux positifs que nos procédures.
La figure 4.3 montre également les taux de détection obtenus avec les knockoffs de Barber et Candès dans le modèle de régression linéaire gaussienne. Pour effectuer leur procédure, il faut choisir une valeur cible pour *false discovery rate*. On veut que cette valeur soit petite mais de trop petites valeurs pour ce FDR cible rendent le seuil infini et donc un modèle estimé vide. Par défaut, le FDR vaut 0.1 mais cette valeur donnait un modèle vide sur trop de répétitions. On

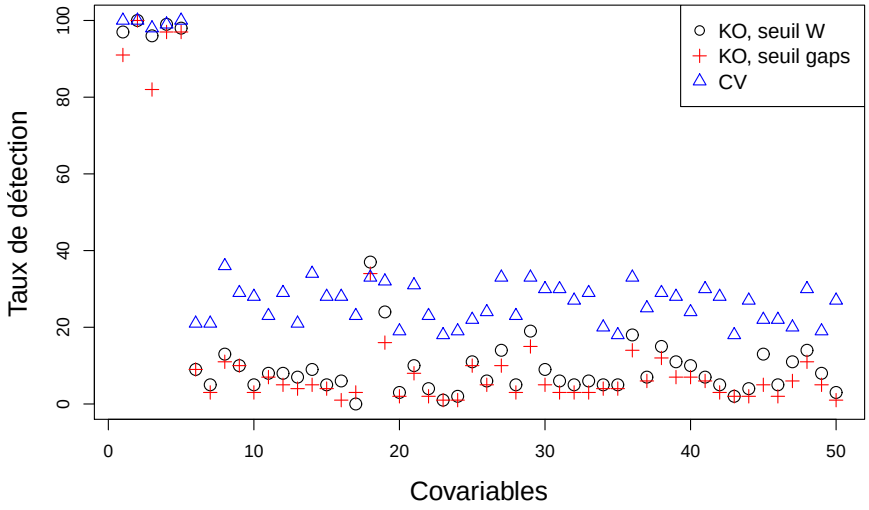


(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$.

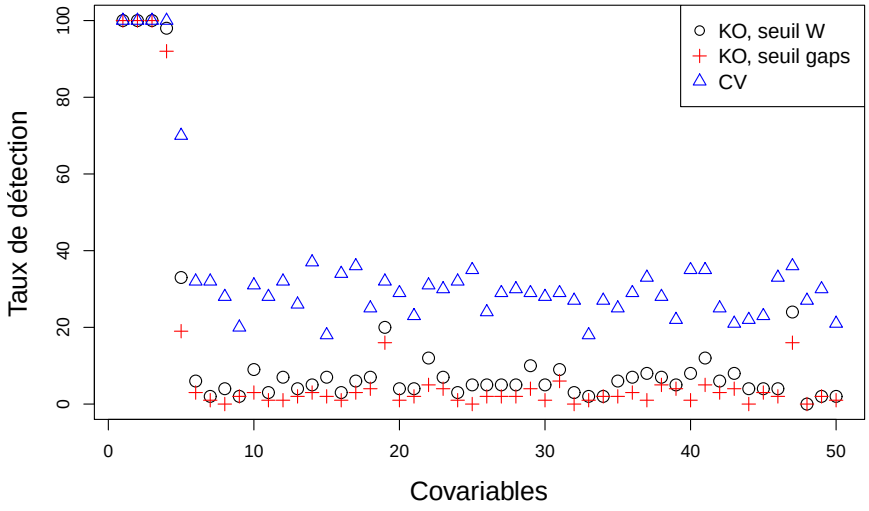


(b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

FIGURE 4.3 – Taux de détection pour chacune des quatre méthodes : les knockoffs revisités avec les seuils W et gaps, la validation croisée (CV) et les knockoffs de Barber et Candès. Régression linéaire gaussienne avec $n = 200$ observations de $p = 50$ covariables. Coefficients de régression $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) et $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.

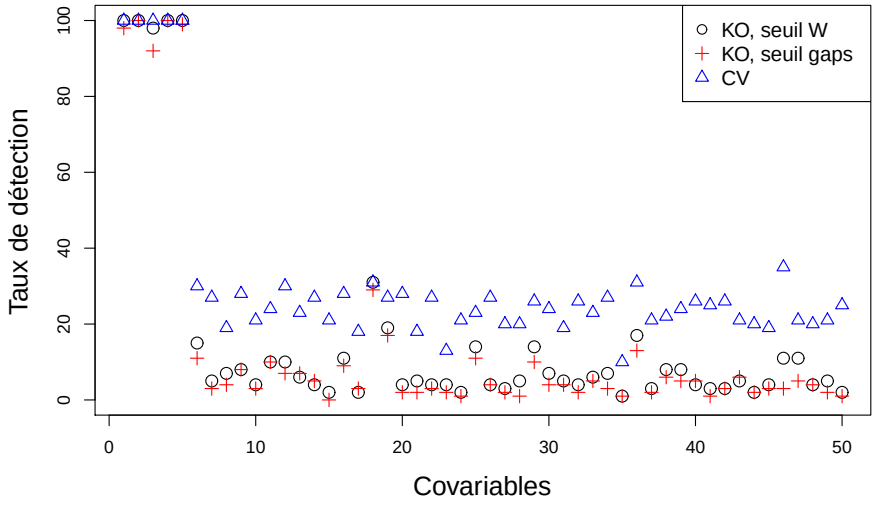


(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$.

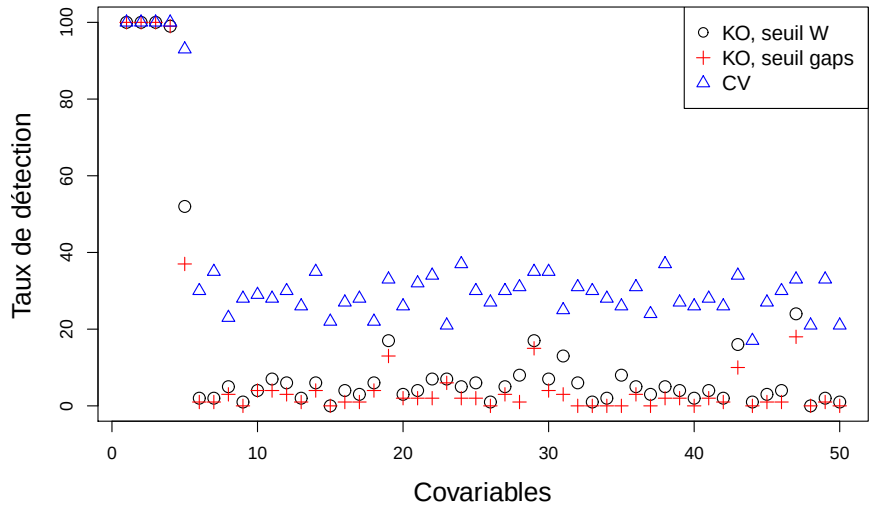


(b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

FIGURE 4.4 – Taux de détection pour chacune des trois méthodes : les knockoffs revisités avec les seuils W et gaps et la validation croisée (CV). Régression logistique avec $n = 200$ observations de $p = 50$ covariables. Coefficients de régression $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) et $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.



(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$.



(b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

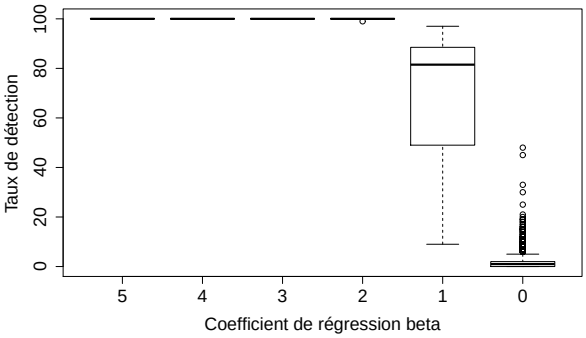
FIGURE 4.5 – Taux de détection pour chacune des trois méthodes : les knockoffs revisités avec les seuils W et gaps et la validation croisée (CV). Régression à logits cumulatifs avec $n = 200$ observations de $p = 50$ covariables. Coefficients de régression $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) et $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.

l'a donc changé à 0.4 pour limiter tous ces modèles estimés vides. Pour les deux configurations différentes de $\boldsymbol{\beta}$, on a obtenu 4 modèles estimés vides sur les 100 répétitions. De ce fait, leur méthode donne des résultats un peu moins bons. En effet, comme le FDR est assez élevé, les taux de détection des covariables nulles ont tendance à être un peu supérieurs aux nôtres. Pour la même raison, les cinq premières covariables sont un peu moins détectées (les taux de détection sont proches de 96%, ce qui correspond au nombre de répétitions pour lequel le seuil n'était pas infini et donc le modèle estimé non vide). Pour $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \dots, 0)$, X_5 est mieux détectée avec leur méthode.
Ces trois figures illustrent aussi que les taux de détection dépendent du coefficient de régression β_k : plus β_k est grand, plus la covariable associée X_k est détectée. En effet, pour $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \dots, 0)$, on peut observer que la covariable X_5 a tendance à être moins détectée que les quatre premières. Par ailleurs, on peut aussi remarquer que certaines des covariables nulles sont plus détectées que d'autres. C'est par exemple le cas pour les covariables nulles $X_{18}, X_{19}, X_{29}, X_{36}, X_{39}$ ou X_{47} pour $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \dots, 0)$ et pour tous les types de régressions (ce qui est cohérent, puisque la matrice design est la même). Ceci est probablement dû à la structure de dépendance du vecteur $\overrightarrow{X}$ des covariables. En particulier, ces covariables sont conditionnellement dépendantes à au moins trois des cinq premières covariables. On observe le même phénomène pour $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \dots, 0)$ avec notamment X_{19}, X_{41} et X_{47} ; les covariables sont alors différentes du cas $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \dots, 0)$ car la matrice design, et donc la structure de dépendance du vecteur $\overrightarrow{X}$, est différente.
Finalement, nos procédures semblent être relativement efficaces, et ce indépendamment du modèle de régression. Les résultats sont néanmoins un peu meilleurs pour la régression linéaire gaussienne. Dans ce cas, on rappelle que la procédure de Barber et Candès [7] offre de meilleures garanties théoriques.

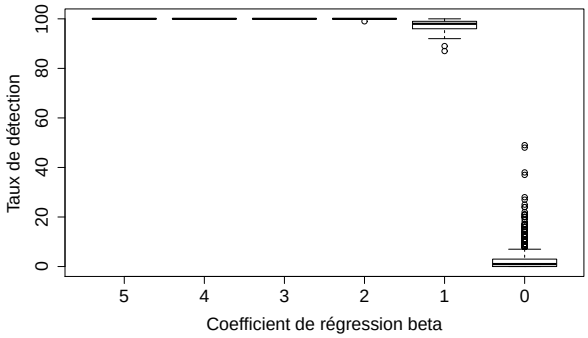
4.2.3 Efficacité et comparaisons - $p = 2000$

On expose à présent des résultats pour $n = 1000$ observations de $p = 2000$ covariables afin d'illustrer les performances de nos procédures sur des milliers de covariables. Les coefficients de régression valent $\beta_k = \begin{cases} 5, \text{ si } 1 \leq k \leq 20, \\ 4, \text{ si } 21 \leq k \leq 40, \\ 3, \text{ si } 41 \leq k \leq 60, \\ 2, \text{ si } 61 \leq k \leq 80, \\ 1, \text{ si } 81 \leq k \leq 100, \\ 0, \text{ sinon.} \end{cases}$. De la même façon que pour $p = 50$ (sous-section 4.2.2), la matrice des covariables $\mathbf{X}$ est la même pour chacune des différentes régressions mais elle diffère selon le vecteur $\boldsymbol{\beta}$ des coefficients de régression. Ainsi, pour une valeur fixée de $\boldsymbol{\beta}$, on a gardé la même matrice design $\mathbf{X}$ pour simuler chacune des trois régressions et la variable réponse $\mathbf{Y}$ est ensuite simulée selon le modèle de régression.

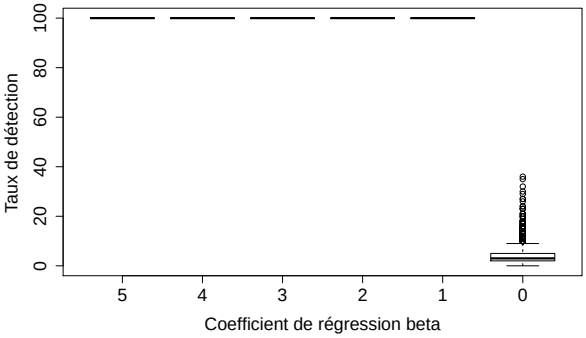
Résultats et commentaires. Les figures 4.6, 4.7 et 4.8 comportent chacune quatre graphiques : les trois premiers sont des boxplots des taux de détection des six groupes de covariables en fonction de leur coefficient de régression β_k. Ces taux de détection sont respectivement obtenus avec les deux méthodes des knockoffs revisités et la validation croisée. Pour mieux comparer notre méthode à la validation croisée pour les covariables nulles (pour lesquelles $\beta_k = 0$), on présente, dans le dernier graphique, les taux de détection des covariables nulles obtenus avec la méthode des knockoffs revisités avec le seuil gaps en fonction des taux de détection obtenus avec la validation croisée. Sur ce graphique, on trace la première diagonale en rouge : ainsi, les points



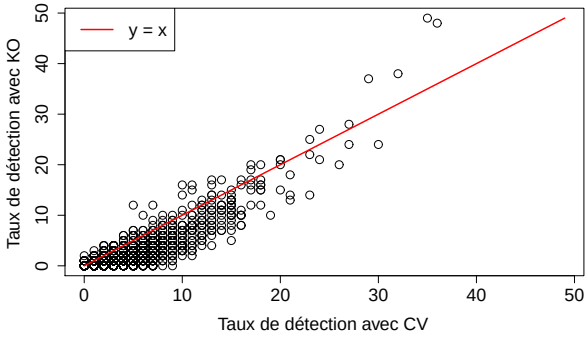
(a) Knockoffs revisités avec le seuil W.



(b) Knockoffs revisités avec le seuil gaps.



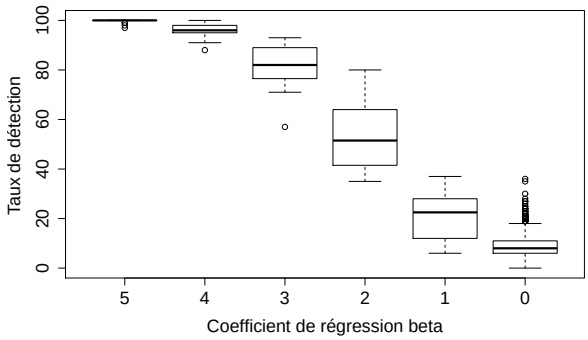
(c) Validation croisée.



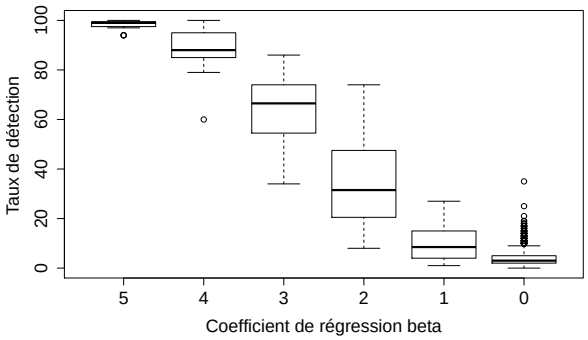
(d) Comparaison des taux de détection des covariables nulles pour la méthode des knockoffs revisités avec le seuil gaps et pour la validation croisée.

FIGURE 4.6 – Boxplots des taux de détection de chacune des covariables en fonction de leur coefficient de régression pour les trois méthodes : les knockoffs revisités avec les seuils W (a) et gaps (b) et la validation croisée (c). Régression linéaire gaussienne avec $n = 1000$ observations de $p = 2000$ covariables. Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.

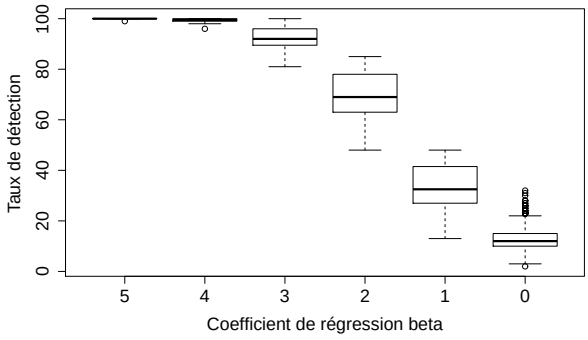
au-dessus correspondent aux covariables nulles qui sont plus détectées avec les knockoffs, ceux qui sont en-dessous aux covariables nulles qui sont plus détectées avec la validation croisée.
Les résultats pour la régression linéaire sont présentés dans la figure 4.6. Au regard des trois boxplots, on remarque que les taux de détection avec les knockoffs revisités avec le seuil W sont plus faibles qu'avec le seuil gaps. Plus précisément, les covariables importantes pour lesquelles $\beta_k = 1$ sont détectées entre 10 et 95% avec le seuil W alors qu'elles sont détectées plus de 80% avec le seuil gaps. Les taux de détection avec le seuil W sont inférieurs et plus dispersés pour les covariables dont le coefficient de régression $\beta_k = 1$ en comparaison aux autres covariables importantes. La validation croisée produit de meilleurs taux de détection pour les covariables importantes. Cependant, la figure 4.6d illustre que la plupart des covariables nulles ont des taux de détection plus élevés avec la validation croisée qu'avec notre procédure. Ainsi, la validation croisée semble donner plus de faux positifs.
Les résultats pour les régressions logistique et à logits cumulatifs sont respectivement présentés dans les figures 4.7 et 4.8. Comme pour $p = 50$, on peut constater sur les boxplots que les taux



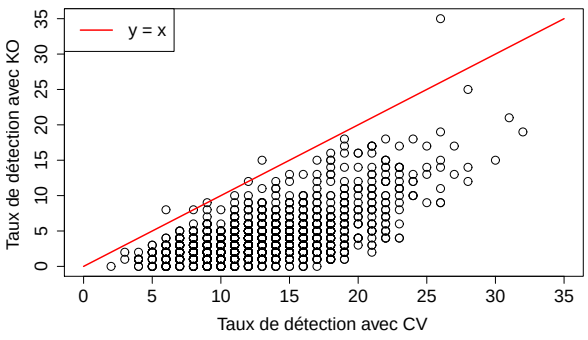
(a) Knockoffs revisités avec le seuil W.



(b) Knockoffs revisités avec le seuil gaps.

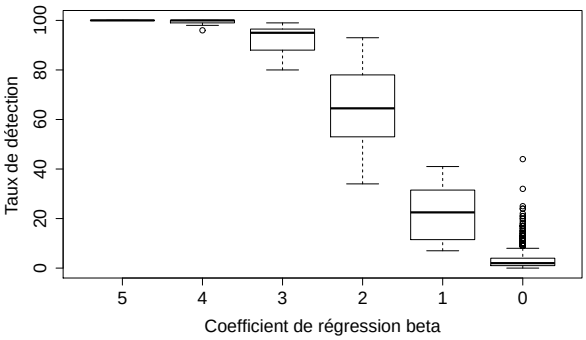


(c) Validation croisée.

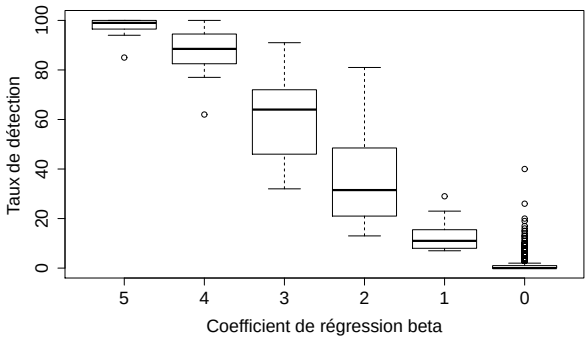


(d) Comparaison des taux de détection des covariables nulles pour la méthode des knockoffs revisités avec le seuil gaps et pour la validation croisée.

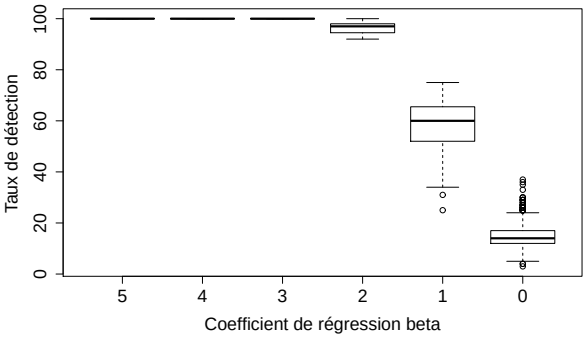
FIGURE 4.7 – Boxplots des taux de détection de chacune des covariables en fonction de leur coefficient de régression pour les trois méthodes : les knockoffs revisités avec les seuils W (a) et gaps (b) et la validation croisée (c). Régression logistique avec $n = 1000$ observations de $p = 2000$ covariables. Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.



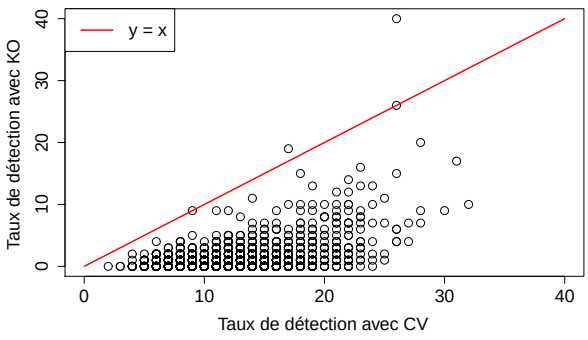
(a) Knockoffs revisités avec le seuil W.



(b) Knockoffs revisités avec le seuil gaps.



(c) Validation croisée.



(d) Comparaison des taux de détection des covariables nulles pour la méthode des knockoffs revisités avec le seuil gaps et pour la validation croisée.

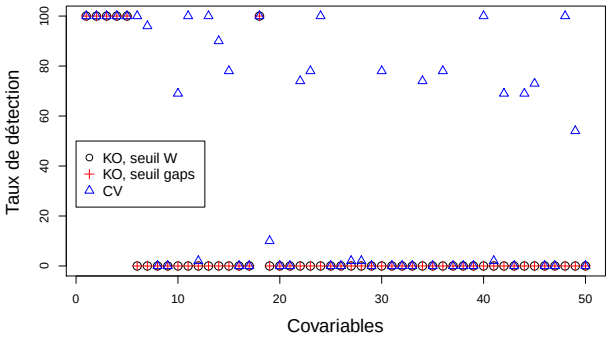
FIGURE 4.8 – Boxplots des taux de détection de chacune des covariables en fonction de leur coefficient de régression pour les trois méthodes : les knockoffs revisités avec les seuils W (a) et gaps (b) et la validation croisée (c). Régression à logits cumulatifs avec $n = 1000$ observations de $p = 2000$ covariables. Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire. Le nombre de répétitions i.i.d. est $B = 100$.

de détection dépendent du coefficient de régression β_k : pour les trois méthodes, les taux de détection sont en effet décroissants selon β_k. Plus β_k est grand, plus les covariables associées sont détectées. On peut aussi observer ce phénomène sur les graphiques 4.6a et 4.6b pour la régression linéaire, bien qu'il soit moins prononcé. Comme pour la régression linéaire, même si la validation croisée détecte mieux les covariables importantes, elle donne également plus de faux positifs parmi les covariables nulles comme on peut le voir sur les graphiques 4.7d et 4.8d. Ce phénomène est encore plus fort pour ces deux modèles de régression pour lesquels presque toutes les covariables nulles sont plus détectées avec la validation croisée qu'avec notre procédure. Contrairement à la régression linéaire, les taux de détection obtenus avec la méthode des knockoffs revisités avec le seuil W sont plus élevés qu'avec le seuil gaps, et ce pour toutes les covariables (importantes comme nulles).
Nos procédures fournissent des résultats satisfaisants pour les trois modèles de régressions présentés, bien que les taux de détection soient meilleurs pour la régression linéaire. Même si les covariables importantes ne sont pas toujours assez détectées, les taux de détection des covariables nulles sont aussi souvent très faibles, particulièrement en comparaison à la validation croisée. De manière générale, notre procédure semble appropriée pour les modèles parcimonieux peu importe le modèle de régression et particulièrement quand le but est d'éviter les faux positifs. Cependant, il faut garder à l'esprit que le seuil à utiliser (seuil W ou gaps) en pratique pour limiter le nombre de faux positifs peut être différent selon le modèle de régression.

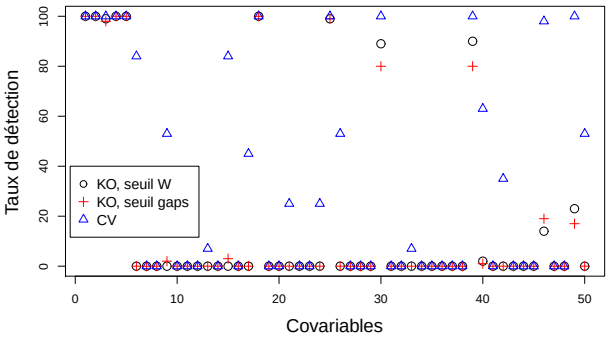
4.2.4 Caractère aléatoire de la procédure

Les procédures des knockoffs revisités et de validation croisée sont toutes deux aléatoires (ce qui n'est pas le cas des knockoffs de Barber et Candès). En effet, notre méthode des knockoffs est aléatoire par la construction de la matrice des knockoffs et la validation croisée dans le découpage de l'échantillon en sous-échantillons. Ainsi, on obtient des résultats différents en appliquant plusieurs fois une de ces méthodes. Pour conclure cette section, on va comparer les taux de détection obtenus par ces trois méthodes (knockoffs revisités avec les deux seuils et validation croisée) sur le même échantillon de données. Cet échantillon est constitué de $n = 200$ observations de $p = 50$ covariables et correspond en fait à un des échantillons parmi les $B = 100$ échantillons utilisés dans la sous-section 4.2.2. La structure de dépendance du vecteur $\overrightarrow{X}$ des covariables est donc la même que dans la sous-section 4.2.2.

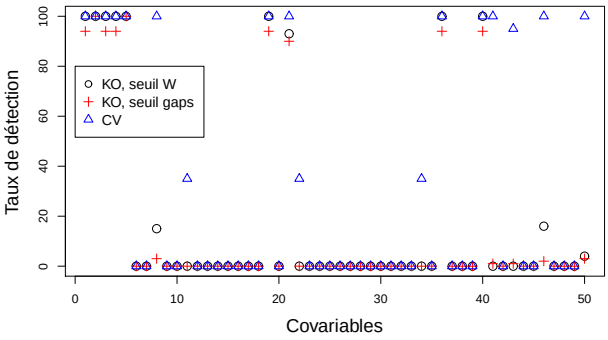
Résultats et commentaires. La figure 4.9 représente les taux de détection de chaque covariable en utilisant le caractère aléatoire des trois procédures : méthode des knockoffs revisités avec les deux seuils (W et gaps) et validation croisée. Pour chacune des trois méthodes, les taux de détection sont obtenus sur 100 répétitions de la méthode sur le même échantillon. On présente les résultats pour les trois modèles de régressions : régressions linéaire, logistique et à logits cumulatifs. Les coefficients de régression valent $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. Ainsi, seules les cinq premières covariables appartiennent au modèle. On peut remarquer que ces cinq premières covariables sont presque toujours détectées à 100% sauf pour le seuil gaps pour la régression à logits cumulatifs (pour laquelle elles sont tout de même détectées à plus de 95%). Les covariables nulles, c'est-à-dire les covariables $X_6, \ldots, X_{50}$, sont là encore moins souvent détectées par nos procédures que par la validation croisée. Cependant, certaines d'entre elles sont détectées à tort avec un pourcentage élevé : X_{18} pour la régression linéaire, X_{18}, X_{25}, X_{30} et X_{39} pour la régression logistique et X_{19}, X_{21}, X_{36} et X_{40} pour la régression à logits cumulatifs. Ceci est probablement lié à l'échantillon, dans lequel la structure de dépendance entre les covariables est plus ou moins prononcée. Rappelons que cette structure de dépendance est la même qu'à la sous-



(a) Régression linéaire.



(b) Régression logistique.



(c) Régression à logits cumulatifs.

FIGURE 4.9 – Taux de détection de chacune des covariables pour les trois méthodes : les knockoffs revisités avec les seuils W et gaps et la validation croisée (CV). 100 répétitions de chacune des trois méthodes sur le même échantillon de $n = 200$ observations de $p = 50$ covariables. Coefficients de régression $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. Les covariables sont des gaussiennes conditionnellement dépendantes avec une structure aléatoire.

section 4.2.2 pour $p = 50$. En comparant aux figures 4.3, 4.4 et 4.5 de la sous-section 4.2.2, on peut constater que ces mêmes covariables ont aussi des taux de détection plus élevés. Pour toutes ces covariables, la validation croisée donne également des taux de détection au moins aussi élevés. Elle détecte aussi beaucoup d'autres covariables nulles. Par exemple, $X_6, X_{11}, X_{13}, X_{24}, X_{40}$ et X_{48} sont toujours détectées (100%) pour la régression linéaire (graphique 4.9a) alors que les knockoffs revisités avec les deux seuils ne les détectent jamais. Pour le modèle à logits cumulatifs, X_8, X_{41}, X_{46} et X_{50} sont toujours détectées alors que nos procédures les détectent moins de 15%.

4.3 Conclusions

Dans ce chapitre, on a exposé une méthode pour la sélection de variables dans des modèles de régressions (impliquant une dépendance linéaire en les covariables) basée sur la construction d'une matrice de copies des covariables. Il s'agit d'une méthode relativement intuitive, qui convient pour un spectre large de régressions, y compris quand le nombre d'observations n est beaucoup plus faible que le nombre p de covariables. On peut choisir deux seuils, ce qui produit deux procédures différentes, qui ont toutes deux été implémentées dans le package R `kosel`. Ces deux procédures s'avèrent pertinentes et efficaces au vu des différents résultats obtenus dans les nombreuses simulations. Nos procédures semblent particulièrement appropriées quand on cherche à éviter les faux positifs. En effet, même s'il y a quelques faux négatifs, il y a surtout un très faible taux de faux positifs. Les simulations montrent aussi que l'efficacité de nos procédures dépend du modèle de régression. D'une manière générale, on conseille d'essayer les deux seuils et de choisir en fonction de son objectif. De plus, le caractère aléatoire de nos procédures ouvre la porte à d'autres techniques pour effectuer la sélection de variables en pratique. Comme proposé à la sous-section 4.2.4, on peut par exemple appliquer la procédure un certain nombre N de fois sur l'échantillon dont on dispose et garder les covariables détectées au-delà d'un certain seuil γN, $0 \leq \gamma \leq 1$, à choisir.
Dans le modèle de régression linéaire gaussienne, on rappelle que la procédure de Barber et Candès est plus sophistiquée et fournit des garanties théoriques. Elle ne convient toutefois qu'à ce modèle et requiert de plus que le nombre d'observations n soit supérieur au nombre p de covariables.
Par ailleurs, nos méthodes donnent de meilleurs résultats que la validation croisée en terme de taux de faux positifs, même dans l'exploitation du caractère aléatoire de ces procédures. Cependant, si on préfère sursélectionner, d'autres techniques, comme la validation croisée, donnent des résultats plus pertinents.

Pour des raisons similaires à celles données dans le chapitre précédent, nous n'avons pas travaillé sur des résultats théoriques mais ceci pourrait faire l'objet de futurs travaux. Nous sommes cependant sceptiques sur l'obtention de résultats généraux indépendants du modèle de régression et il faudrait alors certainement pour cela distinguer les méthodes selon le modèle.

Chapitre 5

Application de modèles de régression pour variables ordinales et de la méthode des knockoffs revisités à l'inférence de réseaux pour données inflatées en zéro

Sommaire

5.1 Modèles pour la simulation de données inflatées en zéro 86
5.1.1 Données gaussiennes inflatées en zéro par des Bernoullis 87
5.1.2 Données *adjacent* 87
5.2 Méthode d'inférence 89
5.3 Résultats 90
5.3.1 Données gaussiennes inflatées en zéro par des Bernoullis 90
5.3.2 Données *adjacent* 92
5.4 Comparaisons avec d'autres méthodes 92
5.4.1 Données gaussiennes inflatées en zéro par des Bernoullis 94
5.4.2 Données *adjacent* 95
5.5 Application à des données réelles 97
5.6 Conclusions 99

On a vu précédemment que plusieurs auteurs avaient utilisé des régressions pour les problèmes d'inférence de réseaux de dépendances conditionnelles en estimant les voisinages de chaque variable. Concernant le modèle graphique gaussien, on a vu dans le chapitre 1 que la nullité d'un coefficient de la régression linéaire est équivalente à la nullité de la corrélation partielle associée et donc à l'indépendance conditionnelle des variables correspondantes. Meinshausen et Bühlmann [71] ont utilisé la régression linéaire pour l'inférence de réseaux dans le modèle graphique gaussien. Même si l'équivalence entre la nullité des coefficients de régression et l'indépendance conditionnelle n'est pas vraie dans tous les modèles, le modèle d'Ising présente également des propriétés similaires. En effet, on a expliqué dans le chapitre 2 que la nullité d'un coefficient de la régression logistique était équivalente à l'indépendance conditionnelle des variables relatives. On a également vu que cette propriété était utilisée en pratique pour inférer des graphes binaires sous le modèle d'Ising notamment par Wainwright (2007) [100], Ravikumar (2010) [82],

Lee (2007) [58] ou encore Jalali (2011) [51].
En *machine learning*, des méthodes similaires [3, 97] sont utilisées pour inférer des réseaux bayésiens, consistant à estimer le voisinage, alors appelé *Markov blanket*, de chacun des sommets. Mais ceci n'est pas développé dans ce manuscrit.

Par ailleurs, l'inférence de réseaux pour données inflatées en zéro, c'est-à-dire pour des données comportant une forte proportion de zéros comme les données d'abondance, ont beaucoup retenu l'attention dernièrement. Par exemple, en biostatistiques où de telles données sont très présentes, Weiss et ses coauteurs (2016) [105] ont recensé et comparé les méthodes existantes de détection basées sur des corrélations simples. Ces méthodes sont ensuite utilisées pour inférer un réseau de dépendances (simples). Dans un contexte de microbiologie, Jakuschkin (2016) [50] a utilisé des méthodes existantes pour l'inférence d'un réseau de micro-organismes végétaux et très récemment, Chiquet et ses coauteurs (2018) [17] ont développé un modèle basé sur un modèle multivarié Poisson log-normal pour l'inférence de réseaux sur des données de comptage.

Dans ce chapitre, on cherche à appliquer le modèle de régression à logits cumulatifs (détaillé dans le chapitre 3) avec la méthode des knockoffs revisités pour la sélection de variables (exposée dans le chapitre 4) à l'inférence de réseaux pour données inflatées en zéro. L'idée consiste à effectuer la régression de chaque variable sur les variables restantes et de sélectionner les covariables (parmi les variables restantes) à l'aide de la méthode des knockoffs revisités. Le voisinage estimé de chaque variable est ainsi constitué des covariables sélectionnées.
Malheureusement, la distribution partielle donnée par le modèle de régression cumulative (3.1) n'est pas consistante avec une distribution jointe (Suggala *et al.* (2017) [93]). Plus précisément, ceci signifie qu'il est difficile voire impossible de simuler un $p-$vecteur X tel que chacune des lois conditionnelles $X_i|X_{-i},\ i = 1, \dots, p$ suive un modèle de régression comme (3.1), la loi jointe n'existant pas en général. Ceci est en fait possible pour des valeurs très particulières de coefficients $\boldsymbol{\beta}$. La régression adjacente, similaire à la régression cumulative, est, elle, en revanche compatible avec une loi jointe.

Par curiosité et manque d'autres méthodes, on souhaite tout de même essayer d'appliquer la méthode des knockoffs pour la régression pénalisée à logits cumulatifs dans un cadre d'inférence de réseaux pour des données inflatées en zéro. Pour ce faire, on va dans un premier temps proposer deux modèles pour simuler des données inflatées en zéros : un modèle gaussien inflaté en zéro par des variables de Bernoullis et le modèle dont les lois conditionnelles suivent le modèle de régression adjacente. Par la suite, on souhaite inférer le réseau en estimant les voisinages de chacune des variables à l'aide de la méthode des knockoffs revisités. On utilisera pour cela les régressions adjacente et à logits cumulatifs.

5.1 Modèles pour la simulation de données inflatées en zéro

Dans cette section, on cherche à simuler des données qui ressemblent à des données d'abondance (positives et naturellement inflatées en zéro) et dont on connaisse la structure de dépendances conditionnelles. Pour cela, on va présenter deux modèles : le premier consiste en un vecteur gaussien qu'on inflate en zéro à l'aide d'un vecteur de Bernoullis ; le second est un modèle tel que chacune des lois conditionnelles suive un modèle de régression adjacente, proche de la régression à logits cumulatifs.

5.1.1 Données gaussiennes inflatées en zéro par des Bernoullis

Ce modèle est un mélange entre un vecteur gaussien, qui spécifie simplement la structure de dépendances conditionnelles, et un vecteur de Bernoullis, pour l'inflation en zéro, dont la loi dépend du vecteur gaussien.

On simule dans un premier temps un $p-$vecteur gaussien X dont la structure de graphe est dite *chain*, c'est-à-dire que X_j et X_k sont dépendantes conditionnellement au reste des variables si et seulement si $|j-k| \leq 1 : X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_p$. On utilise la fonction R `huge.generator` du package `huge` pour la structure de graphe `graph = "band"`. Par défaut, $\mathbb{E}(X_i) = 0$ et $\text{var}(X_i) = 1$ pour tout $i \in \{1, \ldots, p\}$.

Après cela, on change les moyennes et variances de manière à ce que $\mathbb{P}(X_i \geq 0)$ soit proche de 1 pour $i = 1, \ldots, p$ (afin que les données soient positives). Les moyennes $\mu_i,\ i = 1, \ldots, p$ sont choisies entre 1 et 100 : 50% sont choisies uniformément entre 1 et 5, 25% entre 6 et 10, 15% entre 11 et 50 et 10% entre 51 et 100. On fait dépendre les variances σ_i^2 des moyennes de la façon suivante :

$$\sigma_i = \begin{cases} 1.1\frac{\mu_i}{2}, \text{ si } 1 \leq \mu_i \leq 5, \\ 0.9\frac{\mu_i}{2}, \text{ si } 6 \leq \mu_i \leq 10, \\ 0.5\frac{\mu_i}{2}, \text{ si } 11 \leq \mu_i \leq 50, \\ 0.3\frac{\mu_i}{2}, \text{ si } 51 \leq \mu_i \leq 100. \end{cases}$$

Pour finir, on ajoute une inflation en zéro en multipliant le p-vecteur gaussien X par un p-vecteur de Bernoullis Ber, qui dépend du vecteur X. On simule Ber comme suit :

$$\begin{aligned} \text{pour tout } i \in \{1, \ldots, p\},\ Ber_i \sim\ & \mathcal{B}(\pi(X_i)), \\ & \text{où } \pi : \mathbb{R}^+ \to [0,1], \\ & x \mapsto \frac{\exp(a+bx)}{1+\exp(a+bx)}, \text{ où } a = \log(10^{-2}) \text{ et } b = 3. \end{aligned}$$

Ainsi, plus x est proche de 0, plus $\pi(x)$ est proche de 0 et donc, plus la Bernoulli associée est susceptible d'être nulle. Les observations finales sont $Z = Ber \cdot X$. De cette façon, Z_i est plus susceptible d'être nulle si X_i prend des valeurs proches de 0. On dispose de n observations du p-vecteur Z pour inférer le réseau des variables latentes du p-vecteur X dont la structure de dépendances conditionnelles est une chaîne et donnée par la matrice de précision Σ^{-1}.

L'objectif est de retrouver les liens de dépendances conditionnelles entre les variables $X_i,\ i = 1, \ldots, p$, donnés en théorie par la matrice de précision Σ^{-1}, à partir des variables observées $Z_i,\ i = 1, \ldots, p$, inflatées en zéro par le vecteur de Bernoullis, dont la loi dépend de X. Dans ce cas, la structure de graphe sous-jacente est une chaîne, notée par $X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_p$ où $X_i \longleftrightarrow X_j$ représente une arête entre les variables X_i et X_j.

5.1.2 Données *adjacent*

La régression adjacente, *adjacent-categories* ou *consecutive-ratio*, mentionnée brièvement à plusieurs reprises dans cette partie, est une régression pour une variable réponse ordinale. Il s'agit d'un modèle de régression qui convient souvent dans les mêmes situations que la régression à logits cumulatifs [1] et qui possède la même propriété de proportionnalité des odds quand les coefficients de régression ne dépendent pas de la modalité de la variable réponse Y.

Régression adjacente

Soit Y une variable aléatoire à valeurs dans $\{0,\dots,J\}$. Supposons qu'on ait p variables explicatives $X := (X_1, X_2, \dots, X_p)$ et notons $\boldsymbol{\alpha} = (\alpha_0,\dots,\alpha_{J-1}) \in \mathbb{R}^J$, $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p) \in \mathbb{R}^p$ et $\boldsymbol{\beta^*} = (\alpha_0,\dots,\alpha_{J-1},\beta_1,\dots,\beta_p) = (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \mathbb{R}^{J+p}$. Le modèle lie la variable ordinale réponse Y aux variables explicatives $X_1,\dots,X_p$ par les J équations suivantes :

$$\begin{aligned} &\text{logit } \mathbb{P}_{\boldsymbol{\beta^*}}(Y = j | j \le Y \le j+1, X = x) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p \\ \iff \quad &\log\left(\frac{\mathbb{P}_{\boldsymbol{\beta^*}}(Y = j | X = x)}{\mathbb{P}_{\boldsymbol{\beta^*}}(Y = j+1 | X = x)}\right) = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p, \end{aligned} \tag{5.1}$$

pour $j \in \{0,\dots,J-1\}$. Ainsi, la loi de Y sachant $X = x$ est donnée par :

$$\mathbb{P}_{\boldsymbol{\beta^*}}(Y = j | X = x) = \frac{\exp\Big(\sum\limits_{k=j}^{J-1} \eta_k(x)\Big)}{1 + \sum\limits_{l=0}^{J-1} \exp\Big(\sum\limits_{i=l}^{J-1} \eta_i(x)\Big)}$$

pour $j \in \{0,\dots,J\}$ où $\eta_k(x) = \alpha_k + \beta_1 x_1 + \cdots + \beta_p x_p$ et avec la convention $\sum\limits_{k=J}^{J-1} \eta_k(x) = 0$. En d'autres termes, on a $\mathbb{P}_{\boldsymbol{\beta^*}}(Y = J | X = x) = \dfrac{1}{1 + \sum\limits_{l=0}^{J-1} \exp\Big(\sum\limits_{i=l}^{J-1} \eta_i\Big)}$. Ainsi :

$$\mathbb{P}_{\boldsymbol{\beta^*}}(Y | X = x) \propto \exp\Big(\sum_{j=0}^{J-1} \eta_j(x) \mathbb{1}_{Y \le j}\Big),$$

et ce modèle appartient à la famille exponentielle (voir définition 2.1.2).

Loi jointe

À la différence de la régression à logits cumulatifs, cette régression adjacente est consistante avec une distribution jointe (Yang *et al.* (2012) [115]). Plus précisément, soit X un p-vecteur à valeurs dans $\{0,\dots,J\}^p$ dont la loi, qui dépend des paramètres seuils $(\tilde{\theta}_{s,j})_{1\le s\le p, 0\le j\le J-1}$ et de la matrice paramètre symétrique $(\theta_{st})_{1\le s,t\le p}$, est donnée par :

$$\begin{aligned} \mathbb{P}(\mathbf{X}) &\propto \exp\left(\sum_{\substack{1\le s\le p \\ 0\le j\le J-1}} \tilde{\theta}_{s,j} \mathbb{1}_{X_s \le j} + \sum_{1\le s<t\le p} \sum_{j,k=0}^{J-1} \theta_{st} \mathbb{1}_{X_s \le j} \mathbb{1}_{X_t \le k}\right) \\ &\propto \exp\left(\sum_{\substack{1\le s\le p \\ 0\le j\le J-1}} \tilde{\theta}_{s,j} \mathbb{1}_{X_s \le j} + \sum_{1\le s<t\le p} \theta_{st}(J - X_s)(J - X_t)\right) \\ &\propto \exp\left(\sum_{s=1}^{p} \sum_{j=X_s}^{J-1} \tilde{\theta}_{s,j} + \sum_{1\le s<t\le p} \theta_{st}(J - X_s)(J - X_t)\right). \end{aligned}$$

La loi conditionnelle de $X_k|X_{-k}$ est alors :

$$\begin{aligned}
\mathbb{P}(X_k|X_{-k}) &= \frac{\exp\left(\sum_{j=0}^{J-1}\tilde{\theta}_{k,j}\mathbb{1}_{X_k\leq j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\sum_{j=0}^{J-1}\theta_{kt}(J-X_t)\mathbb{1}_{X_k\leq j}\right)}{\sum_{x=0}^{J}\exp\left(\sum_{j=0}^{J-1}\tilde{\theta}_{k,j}\mathbb{1}_{x\leq j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{st}(J-X_t)(J-x)\right)} \\
&= \frac{\exp\left(\sum_{j=0}^{J-1}\tilde{\theta}_{k,j}\mathbb{1}_{X_k\leq j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\sum_{j=0}^{J-1}\theta_{kt}(J-X_t)\mathbb{1}_{X_k\leq j}\right)}{\sum_{x=0}^{J}\exp\left(\sum_{j=x}^{J-1}\tilde{\theta}_{k,j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}(J-X_t)(J-x)\right)} \\
&\propto \exp\left(\sum_{j=0}^{J-1}(\tilde{\theta}_{k,j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}(J-X_t))\mathbb{1}_{X_k\leq j}\right).
\end{aligned}$$

Cette loi conditionnelle suit le modèle de régression adjacente avec :

$$\begin{aligned}
\eta_{k,j}(X_{-k}) &= \tilde{\theta}_{k,j} + \sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}(J-X_t) \\
&= (\tilde{\theta}_{k,j} + J\sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}) - \sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}X_t.
\end{aligned}$$

Par identification, on a : $\alpha_{k,j} = \tilde{\theta}_{k,j} + J\sum_{\substack{1\leq t\leq p\\ t\neq k}}\theta_{kt}$ pour $j = 0,\dots,J-1$ et $\boldsymbol{\beta_k} = (-\theta_{kt})_{t\neq k}$.

La structure de réseau est alors donnée, comme dans le modèle d'Ising, par la matrice $(\theta_{st})_{1\leq s,t\leq p}$: $\theta_{st}\neq 0 \iff X_s \longleftrightarrow X_t$. Comme $\boldsymbol{\beta_s} = (-\theta_{st})_{t\neq s}$, on s'intéressera en pratique à la nullité des coefficients de régression β.

5.2 Méthode d'inférence

Notre approche est la même que celles décrites en introduction de ce chapitre pour les modèles gaussien et d'Ising. Elle consiste à effectuer la régression de chaque variable sur les variables restantes et de sélectionner les covariables (parmi les variables restantes) avec la méthode des knockoffs revisités. De cette façon, on estime les voisinages de chacune des variables dans le réseau. Cette procédure conduit à deux graphes estimés : le graphe "or" et le graphe "and". On dit que la variable X_k appartient au voisinage estimé de la variable X_j si X_k fait partie des covariables sélectionnées par la méthode des knockoffs lorsqu'on effectue la régression de X_j sur les variables restantes $X_1,\dots,X_{j-1},X_{j+1},\dots,X_p$. Le graphe "and" contient une arête entre X_j et X_k si X_k et X_j appartiennent respectivement aux voisinages estimés de X_j et X_k tandis que le graphe "or" contient une arête entre ces deux variables si (au moins) un de ces deux cas se produit. Le graphe estimé "and" est donc inclus dans le graphe estimé "or". On choisit de se restreindre au graphe "and" qui contient moins de faux positifs et qui est donc plus spécifique.

On utilise deux types de régressions : la régression à logits cumulatifs (décrite en (3.1)) et la régression adjacente (décrite en (5.1)). Appliquer la régression adjacente sur les données *adjacent* est cohérent puisque chacune des lois conditionnelles suit ce modèle de régression. Ce n'est en revanche pas le cas pour la régression à logits cumulatifs, ni pour les données gaussiennes inflatées en zéro. D'une part, les régressions adjacente et cumulative sont assez similaires et conviennent aux mêmes types de situations [1] et il est intéressant de comparer les différences que ces régressions donnent en pratique. D'autre part, il est aussi intéressant de voir ce que produisent ces régressions sur un modèle différent pour en étudier la robustesse par exemple.

Pour les données gaussiennes inflatées en zéros par des Bernoullis, il faut toutefois convertir la variable réponse en classes au préalable pour ensuite effectuer la régression sur les variables restantes, qui elles restent inchangées. En effet, que ce soit la régression cumulative ou la régression adjacente, on a besoin de transformer la variable réponse relative pour la rendre ordinale. Par commodité, on note R la variable réponse. R est à ce moment une variable qui comprend un certain nombre de zéros et dont le reste est continu (gaussien). On transforme donc les valeurs non-nulles de R en classes déterminées par les quantiles de R de sorte que ces classes soient équilibrées. Ce type de discrétisation est analogue à ce qui est appelé *equal frequency binning* (Liu (2002) [62]) dans le domaine du *machine learning*. Il reste à choisir le nombre de modalités non-nulles J, qu'on choisit ainsi :

$$J = \left[\frac{\#\{i \in \{1, \dots, n\} \ / \ R^{(i)} \neq 0\}}{20} \right] + 1,$$

où $[.]$ représente la fonction partie entière. Dans nos simulations, J varie de 2 à 11 selon le taux d'inflation en zéro. Quand $J = 2$, la variable discrétisée discrimine seulement si la variable originale vaut 0 ou non.

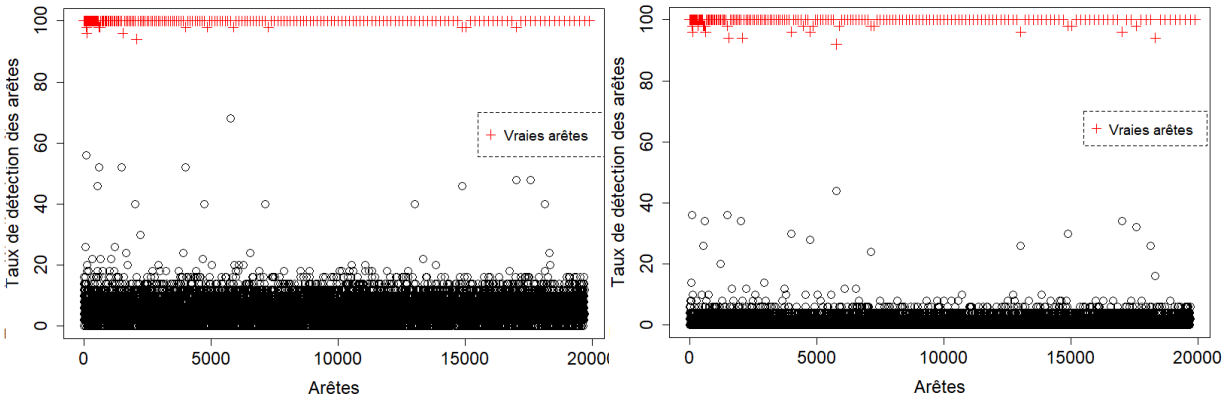
5.3 Résultats

5.3.1 Données gaussiennes inflatées en zéro par des Bernoullis

On simule ici $n = 200$ observations de $p = 200$ variables. Les corrélations entre les variables liées valent environ -0.45 et les corrélations partielles environ -0.38. L'inflation en zéro représente environ 12% en moyenne, variant de 0 à 64% selon les variables. On rappelle que la structure de graphe est une chaîne, notée par $X_1 \longleftrightarrow X_2 \longleftrightarrow \dots \longleftrightarrow X_p$. Le graphe comporte donc 199 arêtes.

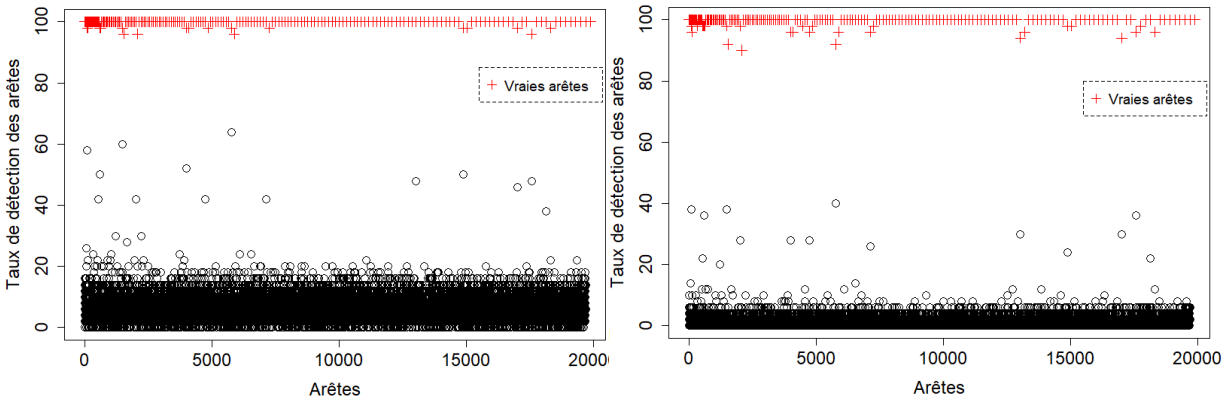
On présente les résultats pour les régressions adjacente et cumulative avec les seuils W et gaps introduits à la sous-section 4.1.3 du chapitre 4.

Résultats et commentaires généraux. Les figures 5.1 et 5.2 représentent les taux de détection de chaque arête sur 50 répétitions indépendantes pour les régressions adjacente et cumulative respectivement sur les données gaussiennes inflatées en zéro. Dans les deux cas, les vraies arêtes sont les plus détectées : toutes sont détectées à plus de 90% et la quasi-totalité d'entre elles sont détectées à 100%. La frontière entre les vraies et les fausses arêtes est relativement distincte, surtout avec le seuil gaps.
Comparaison entre les seuils. Pour les deux types de régression, on peut constater que le seuil gaps donnent de meilleurs résultats : même si certaines des vraies arêtes sont légèrement



(a) Seuil W. (b) Seuil gaps.

FIGURE 5.1 – Taux de détection des arêtes sur 50 répétitions. Méthode des knockoffs revisités pour la régression adjacente. $n = 200$ observations, $p = 200$ variables. Données gaussiennes inflatées en zéro.



(a) Seuil W. (b) Seuil gaps.

FIGURE 5.2 – Taux de détection des arêtes sur 50 répétitions. Méthode des knockoffs revisités pour la régression cumulative. $n = 200$ observations, $p = 200$ variables. Données gaussiennes inflatées en zéro.

moins bien détectées (la différence est de 4% maximum), les fausses arêtes affichent également des taux de détection moindres. En effet, la plupart des fausses arêtes sont détectées entre 0 et 20% avec le seuil W, alors que ces mêmes arêtes sont détectées entre 0 et 12% avec le seuil gaps. Une quinzaine de fausses arêtes sont détectées entre 40 et 65% avec le seuil W tandis qu'elles ne sont détectées qu'entre 20 et 45% avec le seuil gaps.
Comparaison entre les régressions. Les deux types de régression donnent des résultats très similaires. Pour le seuil W, la différence moyenne de taux de détection est de 1.6% avec un écart maximal de 10% ; pour le seuil gaps, la différence moyenne de taux de détection est de 0.4% avec un écart maximal de 6%.

5.3.2 Données *adjacent*

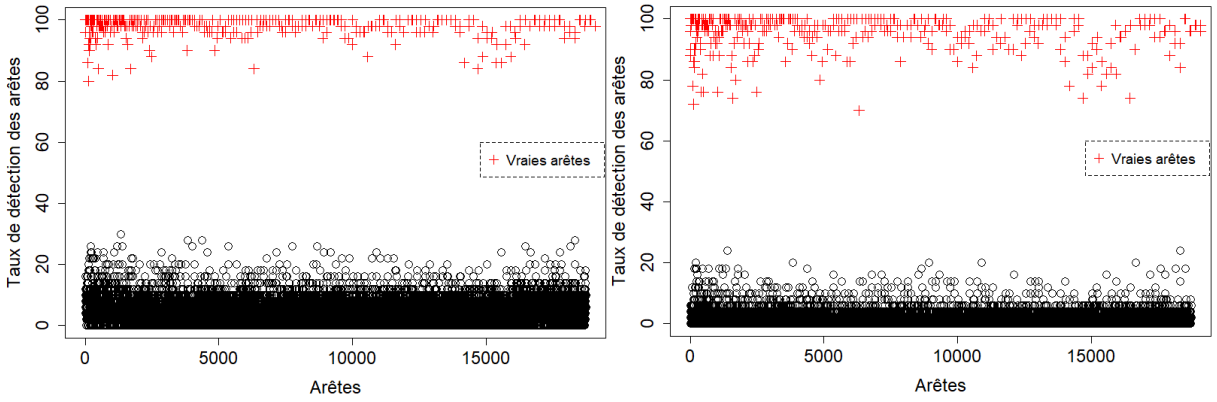
Ici, on simulera des données pour $p = 196 = 14^2$ et $J = 2$. Le réseau a une forme de grille de taille 14×14 et les coefficients de la matrice θ valent ± 0.5 selon si l'arête est verticale ou horizontale. Les paramètres seuils $\tilde{\theta}$ sont choisis uniformément entre -1 et 0 pour essayer d'équilibrer le nombre d'observations pour les modalités de chacune des variables. Les données sont simulées à l'aide de l'échantillonnage de Gibbs décrit à la sous-section 2.2.3 du chapitre 2.

Résultats et commentaires généraux. Les figures 5.3 et 5.4 représentent les taux de détection de chaque arête sur 50 répétitions indépendantes pour les régressions adjacente et cumulative respectivement sur les données *adjacent*. Dans les deux cas, les vraies arêtes sont les plus détectées et la frontière entre les vraies et les fausses arêtes est distincte. Les vraies arêtes sont détectées plus de 80% avec le seuil W et plus de 68% avec le seuil gaps. Les fausses arêtes sont toutes détectées moins de 30% avec le seuil W et moins de 24% avec le seuil gaps.
Assez étonnamment, les vraies arêtes ne sont pas autant détectées que les vraies arêtes des données gaussiennes inflatées en zéro (voir sous-section 5.3.1) alors que le modèle de simulation est adapté aux régressions utilisées (au moins à la régression adjacente en tout cas). En contrepartie, les taux de détection des fausses arêtes sont également un peu plus faibles et la frontière entre les deux types d'arêtes reste nette. Ce phénomène peut être dû à la méthode de simulation de ces données ou aux paramètres choisis. Certaines variables ont peu de valeurs dans une des modalités (selon les paramètres seuils $\tilde{\theta}$ choisis) et ceci impacte la qualité de l'inférence.
Comparaison entre les seuils. De la même façon que pour les données gaussiennes inflatées en zéro, on constate que le seuil gaps a tendance à donner des taux de détection un peu moins importants. Dans les deux régressions, les taux de détection semblent être translatés de quelques pourcents. La translation semble toutefois légèrement plus importante sur les fausses arêtes.
Comparaison entre les régressions. Comme pour les données gaussiennes inflatées en zéro, les deux types de régression donnent ici aussi des résultats très similaires. Pour le seuil W, la différence moyenne de taux de détection est de 1% avec un écart maximal de 8% ; pour le seuil gaps, la différence moyenne de taux de détection est de 0.5% avec un écart maximal de 10%.

5.4 Comparaisons avec d'autres méthodes

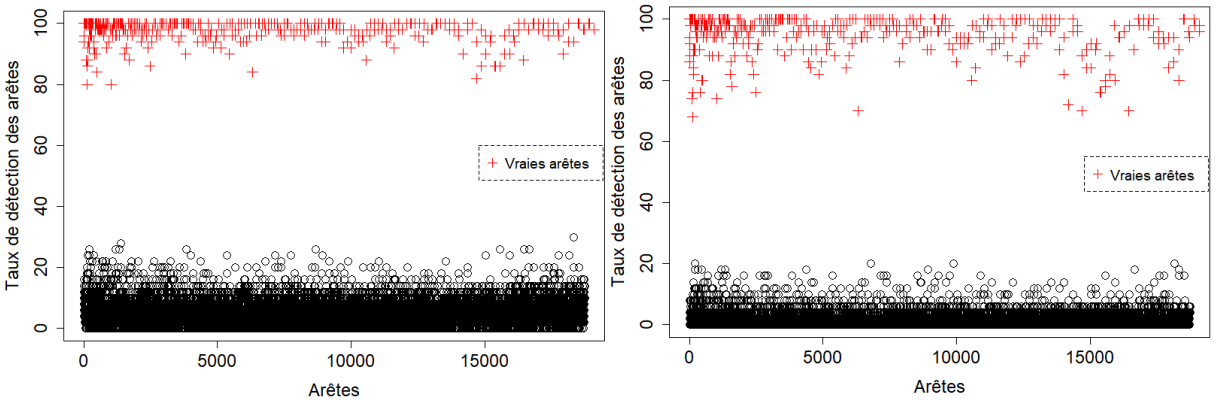
On compare ici les résultats obtenus à la section 5.3 avec les résultats obtenus par quatre autres méthodes.

— Glmnet : La première est la méthode de Meinshausen et Bühlmann [71] concernant le modèle graphique gaussien, qui consiste à estimer le voisinage de chaque variable grâce à une régression linéaire gaussienne L_1-pénalisée. On choisit ensuite d'inférer le graphe



(a) Seuil W. (b) Seuil gaps.

FIGURE 5.3 – Taux de détection des arêtes sur 50 répétitions. Méthode des knockoffs revisités pour la régression adjacente. $n = 200$ observations, $p = 196$ variables. Données *adjacent*.



(a) Seuil W. (b) Seuil gaps.

FIGURE 5.4 – Taux de détection des arêtes sur 50 répétitions. Méthode des knockoffs revisités pour la régression cumulative. $n = 200$ observations, $p = 196$ variables. Données *adjacent*.

"and" décrit ci-dessus. Cette méthode n'est en théorie pas applicable quand les données sont inflatées en zéro, qui contredit l'hypothèse de continuité et donc de normalité des observations. Les paramètres de chacune des régressions linéaires pénalisées sont estimés à l'aide du package R `glmnet` ; le paramètre de pénalisation est choisi par validation croisée à l'aide de la fonction `cv.glmnet`.

— Glasso : La deuxième méthode est la procédure du graphical Lasso de Friedman *et al.* [36] développée dans le cadre du modèle graphique gaussien. Cette procédure consiste à estimer la matrice de précision par maximisation de la log-vraisemblance du modèle gaussien avec une pénalisation Lasso. Les entrées non nulles de la matrice de précision estimée fournissent ensuite le graphe estimé. On effectue cette procédure à l'aide de la fonction `huge`, option `method = 'glasso'` pour le paramètre de pénalisation choisi par le critère "stars".

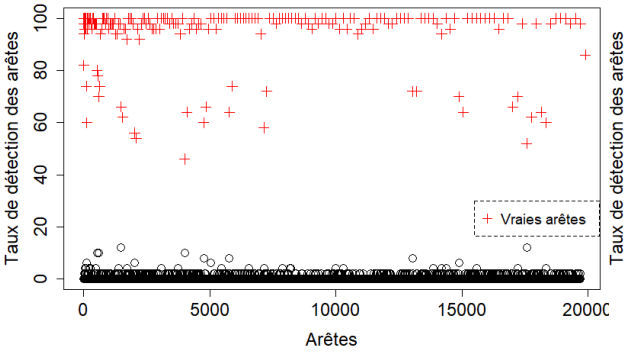
— Pearson et Spearman : Les troisième et quatrième méthodes sont détaillées dans l'article de Weiss *et al.* [105] et reposent sur les corrélations simples de Pearson et Spearman. Pour chacune de ces corrélations, on teste la nullité de chacune des corrélations entre les $\frac{p(p+1)}{2}$ paires de variables sur 1000 échantillons bootstraps. Pour ce faire, on utilise un test basé sur la normalité de la transformation de Fisher : $\rho \mapsto \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ de la corrélation avec un risque $\alpha = 0.05$. On obtient ensuite les p-valeurs empiriques pour chacune des arêtes en regardant $1-$ la proportion de significativité de la corrélation sur les 1000 bootstraps. Ainsi les petites p-valeurs correspondent à une forte proportion de significativité (et donc à une forte probabilité que la corrélation soit non nulle). Enfin, on applique la procédure de correction d'hypothèse multiples de Benjamini-Hochberg aux p-valeurs et les p-valeurs nulles après correction correspondent aux arêtes du réseau estimé. On effectue le calcul des corrélations à l'aide de la fonction `rcorr` du package `Hmisc` et la procédure de correction de Benjamini-Hochberg est effectuée par la fonction `p.adjust`, `method= "BH"`.

Les deux premières méthodes sont basées sur des corrélations partielles (et donc conditionnelles) alors que les deux autres sont basées sur des corrélations usuelles dites simples. Les deux premières méthodes sont adaptées au modèle graphique gaussien et on s'attend à ce qu'elles soient plus pertinentes pour les données gaussiennes inflatées en zéro. Les deux dernières font partie des méthodes souvent utilisées par les biologistes, méthodes que Weiss *et al.* (2016) [105] ont répertoriées et comparées dans un récent article.

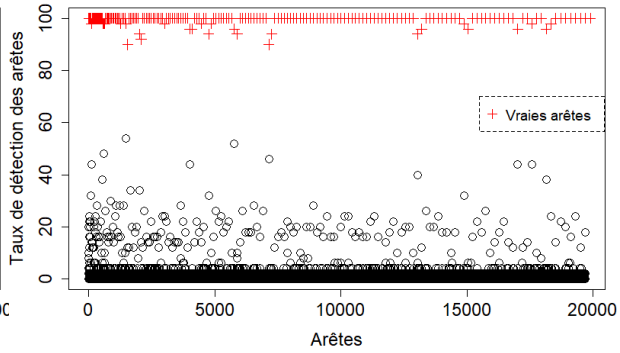
5.4.1 Données gaussiennes inflatées en zéro par des Bernoullis

Résultats et commentaires. La figure 5.5 montre les taux de détection de chaque arête sur 50 répétitions pour les méthodes "Glmnet", "Glasso", "Pearson" et "Spearman" sur les données gaussiennes inflatées en zéro.
En comparaison des résultats obtenus aux figures 5.1 et 5.2, les résultats sont moins bons. La frontière reste assez nette pour Glmnet et Glasso. Pour Glmnet, les fausses arêtes sont moins détectées (moins de 10% environ), mais les vraies arêtes également : la plupart sont détectées plus de 90% mais une trentaine sont détectées entre 40 et 80%. Pour Glasso, les vraies arêtes sont détectées plus de 90%, les fausses arêtes en revanche, sont beaucoup plus détectées (jusqu'à presque 60%).
Pour Pearson et Spearman, les résultats sont assez similaires et la frontière n'est dans les deux cas pas très claire. Les fausses arêtes sont très peu détectées (moins de 10%) et la grande majorité est à 0. En revanche, les vraies arêtes sont mal détectées : une majorité est détectée au-delà de

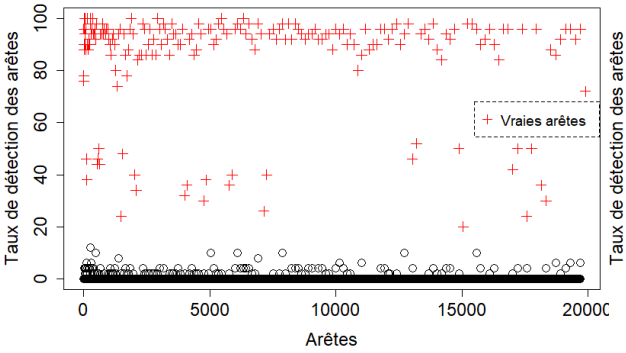
80% mais une trentaine d'arête sont détectées entre 20 et 60%. Pour ces méthodes, qui sont toutes deux basées sur des corrélations simples, on s'attendait plutôt à détecter à tort les arêtes du type $X_i \longleftrightarrow X_{i+2}$, liées indirectement par X_{i+1}. Il est par ailleurs un peu surprenant que les résultats soient aussi proches, dans la mesure où les corrélations de Pearson ont plutôt tendance à déceler des liens linéaires et celles de Spearman des liens monotones, un peu plus généraux.
Les méthodes Glmnet et Glasso, adaptées pour le modèle graphique gaussien, donnent dans ce cas de meilleurs résultats que Pearson et Spearman, ce qui n'est pas incohérent dans la mesure où le modèle des données simulées fait intervenir un modèle gaussien. Les résultats obtenus avec ces deux premières méthodes restent pertinents, même si les résultats des figures 5.1 et 5.2 sont meilleurs.



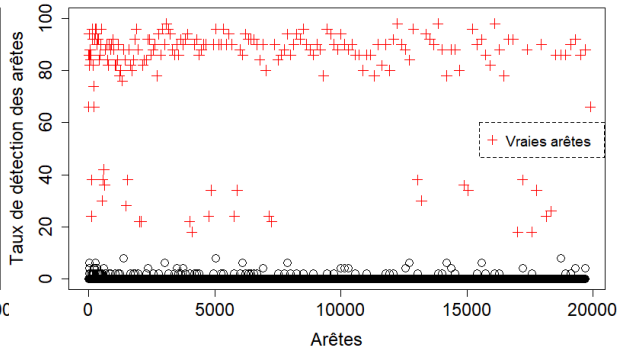
(a) Glmnet.



(b) Glasso.



(c) Pearson.



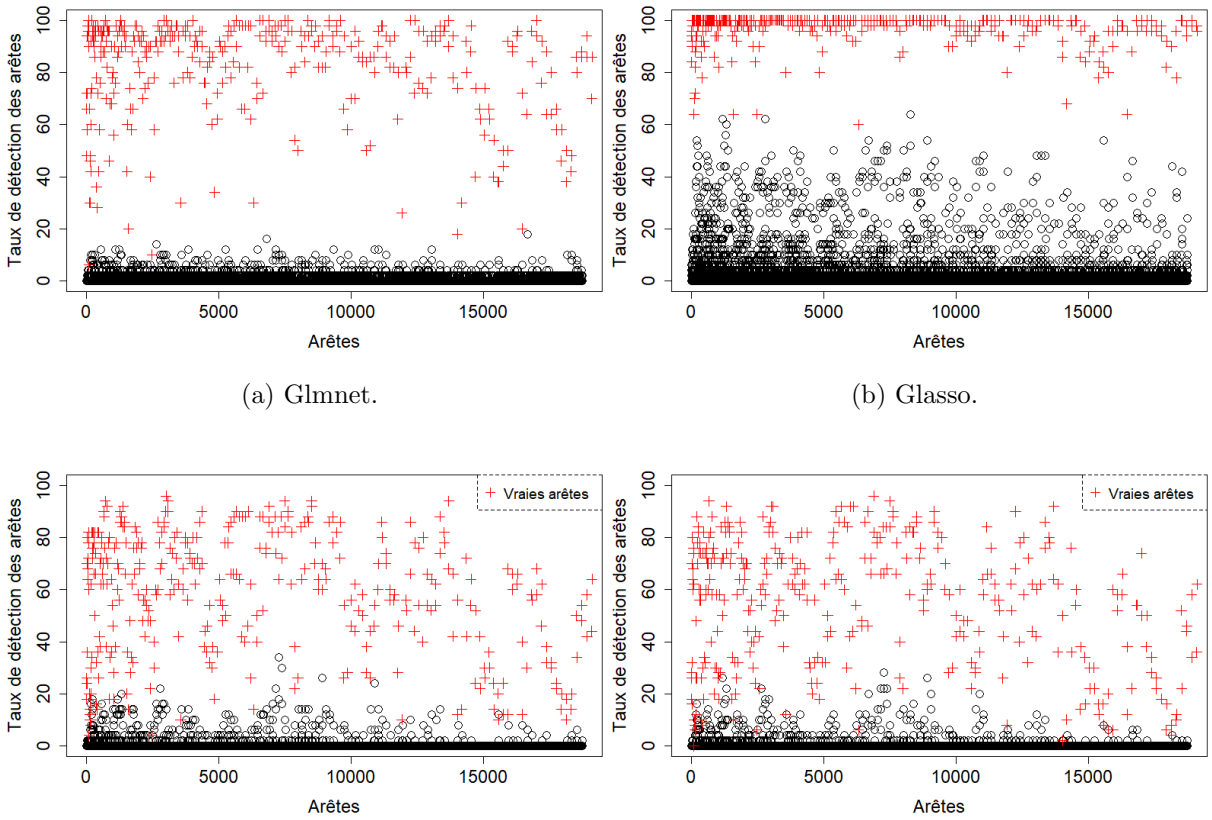
(d) Spearman.

FIGURE 5.5 – Taux de détection des arêtes sur 50 répétitions. Méthodes "Glmnet", "Glasso", "Pearson" et "Spearman". $n = 200$ observations, $p = 200$ variables. Données gaussiennes inflatées en zéro. Les vraies arêtes sont représentées en rouge.

5.4.2 Données *adjacent*

Résultats et commentaires. La figure 5.6 montre les taux de détection de chaque arête sur 50 répétitions pour les méthodes "Glmnet", "Glasso", "Pearson" et "Spearman" sur les données *adjacent*.
Les résultats sont beaucoup plus chaotiques que ceux données par nos procédures (voir figures 5.3

et 5.4), aucune de ces quatre méthodes n'offre une frontière entre les vraies arêtes et les fausses. Les vraies arêtes ont toutefois tendance à être plus détectées que les fausses mais ceci est assez variable d'une méthode à l'autre. Pour Glmnet, les vraies arêtes sont détectées majoritairement au-delà de 60% mais une trentaine d'arêtes sont détectées entre 6 et 60%. Pour Glasso, les vraies arêtes sont détectées plus de 60% mais les fausses arêtes sont détectées jusqu'à 60%, rendant la frontière assez floue. Assez étonnamment, ces deux méthodes donnent de globalement meilleurs résultats que Pearson et Spearman, alors qu'ils sont au départ adaptés au modèle graphique gaussien. Glasso, notamment, estime la matrice de précision du modèle gaussien, ce qui n'a pas de sens ici. Pourtant, c'est la méthode qui donne les meilleures détections des vraies arêtes.
Aucune arête n'est détectée à plus de 94% pour Spearman et Pearson, les détections des vraies arêtes s'étalent entre 4 et 94%. Les fausses arêtes en revanche sont très peu détectées : la plupart à 0% et quelques rares arêtes sont détectées entre 10 et 30%.
Il semble ici plus pertinent d'utiliser une de nos procédures.



(a) Glmnet.

(b) Glasso.

(c) Pearson.

(d) Spearman.

FIGURE 5.6 – Taux de détection des arêtes sur 50 répétitions. Méthodes "Glmnet", "Glasso", "Pearson" et "Spearman". $n = 200$ observations, $p = 196$ variables. Données *adjacent*. Les vraies arêtes sont représentées en rouge.

5.5 Application à des données réelles

Le sujet d'inférence de réseaux pour modèles inflatés en zéro a été initialement motivé par des données de quantités de bactéries se développant autour des truffes. On a donc voulu appliquer cette méthode sur cet ensemble de données, qui a été collecté par des chercheurs de l'INRA (Institut National de la Recherche en Agronomie) et de l'université de Francfort. Ce jeu de données comporte $n = 82$ échantillons de communautés de bactéries se développant autour de la truffe analysées par séquençage à haut débit (données recueillies non-publiées par Splivallo, Vahdatzadeh et Deveau). Un total de 242 OTUs (operational taxonomic units) avec des abondances relatives variant d'un OTU à l'autre ont été obtenus sur les 82 échantillons. Chaque OTU est constitué de groupes de micro-organismes dont les ADN présentent de fortes similarités, c'est-à-dire des micro-organismes très proches génétiquement.

L'objectif de cette étude est de détecter de potentielles interactions entre ces OTUs afin d'identifier de potentiels réseaux d'interactions survenant *in situ* entre ces micro-organismes parmi les milliers d'interactions possibles qui pourraient exister et qui ne peuvent pas être mesurées expérimentalement. Des interactions fonctionnelles peuvent toutefois être analysées expérimentalement sur des petits ensembles de micro-organismes (Faust et Raes (2012) [33]). Ce jeu de données est particulier dans le sens où il contient des données inflatées en zéro : plusieurs OTUs sont en effet présents dans très peu d'échantillons et les données requièrent d'être triées au préalable. Pour cela, on élimine les OTUs présents dans moins d'un tiers des échantillons, c'est-à-dire dans moins de $\frac{82}{3} \approx 27$ échantillons. L'ensemble de données final contient 82 échantillons de $p = 62$ OTUs. Cependant, il reste beaucoup de zéros pour qu'on puisse appliquer les méthodes classiques et performantes du modèle graphique gaussien pour l'analyse des liens entre ces variables. On applique alors notre procédure des knockoffs revisités (seuil gaps) pour la régression à logits cumulatifs.

La procédure des knockoffs est aléatoire par la construction de la matrice des knockoffs (voir sous-section 4.2.4 du chapitre 4). Cet aléa permet de pondérer les arêtes et de les trier en fonction. Dans ce travail, on choisit de répéter 80 fois la procédure des knockoffs revisités et de sélectionner les arêtes qui sont détectées plus de 60 fois sur les 80 répétitions. On pondère ensuite chacune de ces arêtes en fonction de son nombre de détection.

Résultats et commentaires. Notre procédure produit un réseau contenant un total de 50 arêtes entre 44 OTUs et dont la composante connexe principale contient 33 OTUs et 42 arêtes. La figure 5.7 représente cette composante connexe principale. Le réseau est organisé en deux clusters liés par 3 OTUs. Le premier cluster (du haut) est constitué d'OTUs qui ont tendance à fonctionner ensemble et qui correspondent à des bactéries étroitement liées en terme de capacités fonctionnelles. Ces OTUs ont tendance à cohabiter et exister naturellement ensemble et à interagir au sein des truffes. Les OTUs du deuxième cluster (en bas) ont aussi tendance à co-exister alors que les OTUs 1 et 2 qui connectent les deux clusters montrent une tendance d'exclusion. Un lien négatif similaire entre les OTUs 1 et 2 est également observé dans d'autres jeux de données (Splivallo *et al.* (2019) [91], corroborant la validité de ce constat. Un tiers des OTUs impliqués dans ce cluster sont disponibles en culture en laboratoire à l'INRA et des tests expérimentaux pourraient être effectués dans le futur pour valider ces hypothèses.

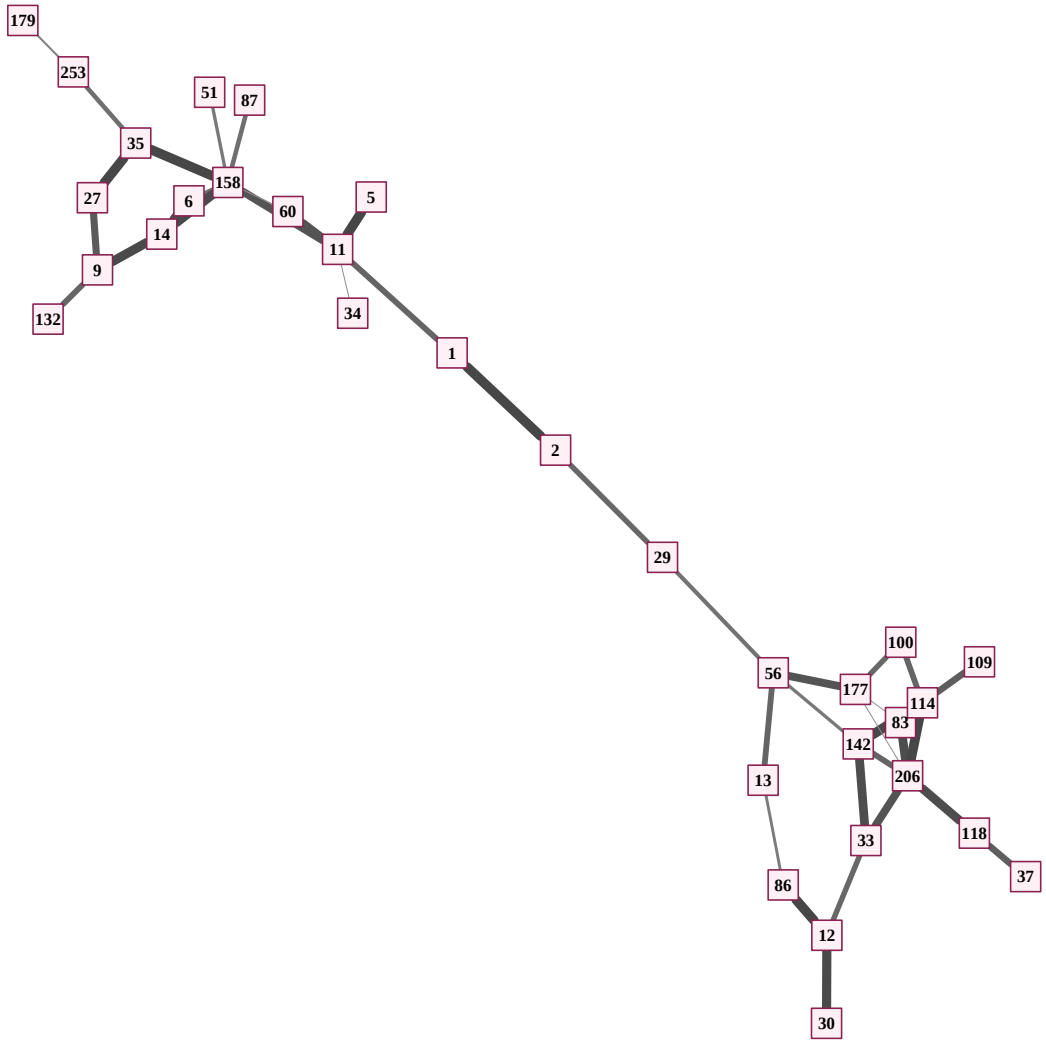


FIGURE 5.7 – Composante connexe principale du réseau des OTUs de la truffe obtenu après 80 applications de la méthode des knockoffs. Les arêtes sont représentées si elles sont détectées plus de 60 fois et elles sont pondérées selon leur nombre de détection.

5.6 Conclusions

Dans ce chapitre, on s'est focalisé sur l'inférence de réseaux dans des modèles inflatés en zéro. L'idée est d'estimer les voisinages de chacune des variables à l'aide de la procédure des knockoffs revisités, détaillée au chapitre 4. On a appliqué cette procédure pour deux modèles de régression : le modèle de régression à logits cumulatifs, présenté au chapitre 3, et un modèle de régression très similaire, le modèle de régression adjacente. On a simulé des données inflatées en zéro de deux façons : avec un modèle de données gaussiennes inflatées en zéro par des variables de Bernoulli et un modèle dont chacune des lois conditionnelles suit le modèle de régression adjacente.

Les résultats obtenus sont relativement bons. Les vraies arêtes semblent même être mieux détectées sur les données gaussiennes inflatées en zéro. Le seuil gaps a tendance à donner des résultats légèrement meilleurs : les vraies arêtes sont un peu moins bien détectées mais les fausses arêtes également, et de façon plus marquée. Le choix de la régression à logits cumulatifs ou de la régression adjacente ne semble pas déterminant, les résultats obtenus avec ces deux régressions sont très proches. Cette constatation a déjà été établie par Agresti (2010) [1] et est liée au fait que ces modèles de régression font tous les deux intervenir des modalités ordonnées pour Y dans la modélisation de sa distribution.
Même si les modèles de régression ne sont pas forcément adaptés aux données (c'est le cas des données gaussiennes), les résultats sont loin d'être incohérents. Ceci donne de l'espoir quant à la robustesse de cette méthode.
De plus, cette procédure a donné des résultats plutôt encourageants sur des données réelles dans un contexte agronomique, certains des liens estimés ayant déjà été validés ou observés par les chercheurs.

Nous ne présentons dans ce chapitre que deux modèles de simulation de données inflatées en zéro. Ces modèles ne sont pas évidents à trouver, car ils font intervenir un certain nombre de contraintes : notamment, la structure de dépendances conditionnelles doit être accessible et il faut pouvoir modéliser l'inflation en zéro. Par la suite, il pourrait être intéressant et pertinent de déterminer d'autres modèles théoriques sur lesquels tester cette procédure.

Troisième partie

Modèles inflatés en zéro

Chapitre 6

Inférence de réseaux pour données gaussiennes inflatées en zéro par double troncature

Sommaire

6.1 Modèle théorique et procédure d'estimation 104
6.1.1 Présentation du modèle 104
6.1.2 Quelques outils théoriques 105
6.1.3 Estimation 106
6.2 Résultats de convergence 109
6.2.1 Estimateurs des points de troncature 109
6.2.2 Estimateur de la matrice de covariance 111
6.2.3 Estimateur de la matrice de précision 116
6.3 Simulations 119
6.3.1 Paramètres de simulations 119
6.3.2 Choix du paramètre de pénalisation 121
6.3.3 Efficacité de la procédure 123
6.3.4 Impact de l'estimation des points de troncature 125
6.3.5 Impact des points de troncature 127
6.3.6 Autres structures de graphes 129
6.4 Conclusions 132

Les modèles de gaussiennes tronquées ont beaucoup attiré l'attention tout au long de la deuxième moitié du siècle dernier. Cohen (1949 [18], 1950 [19], 1957 [22]) a notamment beaucoup étudié l'estimation de la moyenne et de la variance pour des données gaussiennes univariées tronquées de façon unilatérale comme bilatérale (*singly or doubly truncated gaussian*). Dans ces articles, il propose l'estimation de ces paramètres par maximisation de la vraisemblance et montre que ces estimateurs coïncident avec ceux de la méthode des moments (jusqu'à l'ordre 2). Cette procédure conduit à la résolution d'un système à deux inconnues pour la résolution duquel des tables ont été construites et complétées au fil des années. Cohen y distingue également les cas où le nombre d'observations tombant dans les queues de distributions (au-delà des points de troncatures) est connu ou non. Une synthèse de ces résultats est proposée dans le livre de Cohen (1991) [20]. Le cas univarié a également aussi été étudié par Shah et Jaiswal (1966) [88], qui

proposent d'estimer ces paramètres à l'aide des quatre premiers moments empiriques.

Le cas bivarié a par la suite été naturellement étudié. Campbell (1945) [13] étudie le cas d'un vecteur gaussien où les deux variables sont tronquées mais non corrélées. Raj (1953) [81] et Cohen (1955) [21] étudient l'estimation des moyennes, variances et covariance dans le cas où seule une des variables du vecteur est tronquée. Raj distingue les cas où on connaît le nombre d'observations qui tombent en-dehors des points de troncature ou non. Dans le cas où ce nombre est connu, la valeur de la variable non tronquée est inconnue. Par la suite, l'estimation de ces paramètres est proposée dans le cas où les deux variables sont tronquées. Nath (1971) [74] étudie cette estimation par la méthode du maximum de vraisemblance tandis que Dyer (1973) [30], puis Muthen (1990) [73] l'étudient par la méthode des moments. Dans tous ces articles, dès qu'une des variables tombe en-dehors de ses points de troncature, aucune des variables du couple n'est observée.

Concernant le cas multivarié, Cohen (1957) [23] estime les paramètres du modèle où seule une variable du vecteur est tronquée en maximisant la vraisemblance. Singh (1960) [90] s'intéresse à l'estimation des moyennes et variances dans le modèle où seulement certaines variables du vecteur gaussien sont tronquées. Par la suite, Gupta et Tracy (1976) [41], Lee (1983) [57] et Manjunath et Wilhelm (2009) [9] ont étudié les moments dans le cas d'un vecteur gaussien dont toutes les variables sont "doublement" tronquées, c'est-à-dire de façon bilatérale.

Dans ce chapitre, on s'intéresse à l'estimation de la matrice de covariance et de la matrice de précision dans le cas d'un vecteur gaussien dont toutes les variables sont doublement tronquées. Dans notre cas, chacune des variables est observée "normalement" à l'intérieur de ses points de troncature et vaut 0 en-dehors. La double troncature consiste donc à observer un 0 lorsque la gaussienne initiale tombe en-dehors de ses points de troncature associés, mais permet d'observer les autres variables du vecteur, contrairement à ce qui existe à ce sujet dans la littérature. Plus particulièrement, on va chercher à retrouver la structure de graphe du vecteur gaussien, donnée en théorie par la matrice de précision, à partir des observations tronquées. À notre connaissance, aucun résultat de la littérature ne concerne ce problème d'inférence de réseaux dans le cas de gaussiennes inflatées en zéro par troncature.

On donne notamment un théorème qui précise la vitesse de convergence de l'estimateur de la matrice de précision et spécifie des conditions sous lesquelles le graphe théorique est correctement estimé, avec probabilité convergeant vers 1.

6.1 Modèle théorique et procédure d'estimation

6.1.1 Présentation du modèle

Soit X un p-vecteur gaussien $X \sim \mathcal{N}_p(\mu, \Sigma^*)$ où $\mu = (\mu_j)_{j=1,\dots,p} \in \mathbb{R}^p$ est le vecteur des moyennes et $\Sigma^* = (\Sigma^*_{jk})_{1\leq j,k\leq p} \in \mathcal{M}_p(\mathbb{R})$ la matrice de covariance. On considère le vecteur Y défini par $Y_j = \mathbb{1}_{a_j \leq X_j \leq b_j} X_j$ pour tout $j \in \{1, \dots, p\}$ où les points de troncature $a_j, b_j \in \mathbb{R}$ sont tels que $a_j < b_j$ et dépendent de j. Le vecteur gaussien X n'est donc pas observé directement mais observé au travers du vecteur Y tronqué. Contrairement à ce qui existe dans la littérature sur les gaussiennes tronquées, lorsqu'une des variables gaussiennes tombe en-dehors de ses points de troncature, on observe un 0 et on observe le reste du vecteur selon la même règle. Autrement dit, la troncature ne consiste pas à se restreindre aux observations de X qui tombent dans le

pavé $[a_1, b_1] \times \cdots \times [a_p, b_p]$ de $\mathbb{R}^p$.

Quitte à estimer μ_j et Σ^*_{jj} grâce aux méthodes présentées dans la littérature des gaussiennes univariées "doublement" tronquées (voir par exemple, [22]), on se restreint ici au cas où X est centré et réduit, c'est-à-dire $\mu_j = 0$ et $\Sigma^*_{jj} = 1$ pour tout $j \in \{1, \ldots, p\}$.

Rappelons que le modèle gaussien est particulièrement approprié à l'inférence de graphe d'indépendance conditionnelle. Ces dépendances conditionnelles sont en effet spécifiées par la matrice de précision $\Theta^* := (\Sigma^*)^{-1}$ du vecteur gaussien X, qui fournit facilement cette structure latente de graphe. Plus précisément, ce graphe contient une arête entre les variables X_j et X_k si :

$$\begin{aligned} X_j \longleftrightarrow X_k &\iff X_j \not\perp\!\!\!\perp X_k \mid (X_l)_{l \neq j,k} \\ &\iff \mathrm{cor}(X_j, X_k \mid (X_l)_{l \neq j,k}) \neq 0 \\ &\iff \Theta^*_{jk} \neq 0. \end{aligned}$$

Pour plus de détails, on pourra se référer au chapitre 1. L'objectif de ce chapitre est de retrouver la structure latente de graphe, donc la structure de dépendances conditionnelles entre les variables du vecteur X donnée par Θ^*, à partir d'observations du vecteur Y.

6.1.2 Quelques outils théoriques

Afin de davantage expliciter le modèle et d'en exhiber certaines complexités, on va dans un premier temps écrire la vraisemblance d'un couple de variables du vecteur Y.

Soit $(j, k) \in \{1, \ldots, p\}^2,\ j < k$. On note $f_{jk}(x, y)$ la densité du couple gaussien $(X_j, X_k) \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \Sigma^*_{jk} \\ \Sigma^*_{jk} & 1 \end{pmatrix} \right)$. On a $f_{jk}(x, y) = f(x, y, \Sigma^*_{jk})$ où :

$$\begin{aligned} f(x, y, \sigma) &= \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\left[-\frac{1}{2}(x\ y) \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right] \\ &= \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\left[-\frac{x^2 - 2\sigma xy + y^2}{2(1-\sigma^2)} \right]. \end{aligned}$$

La vraisemblance du couple (Y_j, Y_k) est $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ où y correspond à la réalisation du vecteur Y et :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k), \tag{6.1}$$

avec :

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j = a, \zeta_k = b}$ où $\zeta_l = \begin{cases} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{cases}$
- $\sum_{a,b=0}^{1} n_{ab}(y_j, y_k) = 1$
- $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$
- $\phi_{01,jk}(\sigma, y_j, y_k) = \phi_{01,jk}(\sigma, y_k) = \int_{[a_j, b_j]^c} f(x, y_k, \sigma) \mathrm{d}x$

— $\phi_{10,jk}(\sigma, y_j, y_k) = \phi_{10,jk}(\sigma, y_j) = \displaystyle\int_{[a_k,b_k]^c} f(y_j, y, \sigma)\mathrm{d}y$

— $\phi_{00,jk}(\sigma, y_j, y_k) = \phi_{00,jk}(\sigma) = \displaystyle\iint_{[a_j,b_j]^c\times[a_k,b_k]^c} f(x, y, \sigma)\mathrm{d}x\mathrm{d}y.$

On peut ainsi constater que la vraisemblance (et donc la log-vraisemblance) d'un couple de variables de Y fait intervenir quatre termes selon la nullité de chacune des composantes du couple. De la même façon, pour écrire la vraisemblance du vecteur Y, on aurait alors 2^p termes en distinguant tous ces cas. Cette vraisemblance fait intervenir la densité du vecteur gaussien (aucune variable de Y n'est nulle), p intégrales simples (une seule variable nulle), $\binom{p}{2}$ intégrales doubles (deux variables nulles), $\binom{p}{3}$ intégrales triples, ..., une intégrale p-multiple (toutes les variables sont nulles). Écrire la vraisemblance du p-vecteur Y devient alors assez complexe et fastidieux. Par la suite, on choisit donc de se limiter à la vraisemblance des couples pour l'estimation des matrices de covariance et précision.

En pratique, on dispose d'un n-échantillon $\mathbf{Y} := (Y^{(1)}, \ldots, Y^{(n)})$ du vecteur Y. La vraisemblance de l'échantillon $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\ldots,n}$ devient donc $\mathcal{L}_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ définie par :

$$\begin{aligned}\mathcal{L}_{jk}^{(n)}(\sigma, \mathbf{y}) &= \prod_{i=1}^{n} \mathcal{L}_{jk}(\sigma, y^{(i)}),\\ &= \prod_{i=1}^{n}\sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j^{(i)}, y_k^{(i)}) n_{ab}(y_j^{(i)}, y_k^{(i)}),\end{aligned}$$

où $\mathbf{y} := (y^{(1)}, \ldots, y^{(n)})$ correspond à la réalisation du n-échantillon $\mathbf{Y}$. La log-vraisemblance vaut alors $L_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ où :

$$\begin{aligned}L_{jk}^{(n)}(\sigma, \mathbf{y}) &= \sum_{i=1}^{n}\sum_{a,b=0}^{1} n_{ab}(y_j^{(i)}, y_k^{(i)}) \log\left(\phi_{ab,jk}(\sigma, y_j^{(i)}, y_k^{(i)})\right)\\ &= \sum_{\substack{i=1\\ i:y_j^{(i)}=y_k^{(i)}=0}}^{n} \log\left(\phi_{00,jk}(\sigma)\right) + \sum_{\substack{i=1\\ i:y_j^{(i)}=0, y_k^{(i)}\neq 0}}^{n} \log\left(\phi_{01,jk}(\sigma, y_k^{(i)})\right)\\ &+ \sum_{\substack{i=1\\ i:y_j^{(i)}\neq 0, y_k^{(i)}=0}}^{n} \log\left(\phi_{10,jk}(\sigma, y_j^{(i)})\right) + \sum_{\substack{i=1\\ i:y_j^{(i)}\neq 0, y_k^{(i)}\neq 0}}^{n} \log\left(\phi_{11,jk}(\sigma, y_j^{(i)}, y_k^{(i)})\right).\end{aligned}$$

Remarque : Des explorations numériques montrent que $\sigma \mapsto -L_{jk}^{(n)}(\sigma, \mathbf{y})$ n'est pas convexe.

6.1.3 Estimation

Étape 1 : estimation de la matrice de covariance

Une première étape intermédiaire consiste à estimer la matrice de covariance Σ^* du vecteur X à partir d'observations du vecteur Y.

Estimer cette matrice de covariance à partir de la matrice de covariance empirique du n-échantillon $\mathbf{Y}$ conduirait à des résultats médiocres en raison du nombre important de zéros.

Par ailleurs, dans la littérature des gaussiennes tronquées bivariées comme multivariées, les données considérées sont seulement celles dont toutes les composantes tombent entre les points de troncature c'est-à-dire qu'elles correspondraient aux observations du vecteur Y pour lesquelles toutes les composantes sont non nulles. Ces méthodes sont difficilement exploitables dans notre cas car elles réduiraient drastiquement le nombre d'observations, et ceci même dans le cas bivarié si on cherchait à estimer une à une les entrées de la matrice de covariance à l'aide du couple correspondant de variables du vecteur Y.

Une autre idée pourrait être de l'estimer par maximisation de la vraisemblance du vecteur Y. Or, cette vraisemblance est très difficile à écrire. On voit que la vraisemblance d'un couple de variables du vecteur Y, donnée en (6.1), est une somme de quatre termes permettant de traiter les quatre cas possibles (selon la nullité de chacune des variables). Ainsi, la vraisemblance du vecteur Y se découperait en 2^p termes correspondant à des intégrales multiples de la densité du vecteur gaussien X en dehors des points de troncature. Par ailleurs, cette vraisemblance ne serait pas convexe non plus et mènerait probablement à un problème d'optimisation difficile à résoudre.

Au regard de ces difficultés, nous avons décidé d'estimer cette matrice de covariance terme à terme en se ramenant à l'étude des couples (Y_j, Y_k), $j < k$ de variables du vecteur Y. Pour les mêmes raisons qu'expliquées précédemment, la littérature des gaussiennes bivariées tronquées, comme [74], [30] ou encore [73], ne semble pas convenir à ce cas.

Finalement, nous proposons d'estimer Σ^* par $\widetilde{\Sigma}^{(n)}$ en estimant chacun de ses coefficients Σ^*_{jk} par maximisation de la log-vraisemblance $L^{(n)}_{jk}(\sigma, \mathbf{y})$ de l'échantillon $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\dots,n}$ du couple (Y_j, Y_k).

Définition 6.1.1 (Estimateur $\widetilde{\Sigma}^{(n)}$ de Σ^*)**.** *L'estimateur $\widetilde{\Sigma}^{(n)} = (\widetilde{\Sigma}^{(n)}_{jk})_{1\leq j,k\leq p}$ de la matrice de covariance Σ^* est défini par :*

$$\begin{aligned} \widetilde{\Sigma}^{(n)}_{jk} = \widetilde{\Sigma}^{(n)}_{kj} &:= \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ L^{(n)}_{jk}(\sigma, \boldsymbol{y}) \\ &= \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ \frac{1}{n} L^{(n)}_{jk}(\sigma, \boldsymbol{y}), \textit{ pour tout } 1 \leq j < k \leq p, \\ \widetilde{\Sigma}^{(n)}_{jj} &:= 1, \textit{ pour tout } 1 \leq j \leq p. \end{aligned} \tag{6.2}$$

Étape 2 : estimation de la matrice de précision

Notre objectif étant de retrouver le graphe d'indépendance conditionnelle sous-jacent, il est naturel d'utiliser l'estimateur de la matrice de précision Θ^* donné par le graphical Lasso [36], déjà introduit dans la sous-section 1.2.1 du chapitre 1. Le graphical Lasso est une procédure utilisée dans le modèle graphique gaussien qui consiste à estimer la matrice de précision en maximisant la log-vraisemblance pénalisée du modèle gaussien sur l'ensemble des matrices définies positives de dimension $p \times p$:

$$\underset{\Theta \succ 0}{\operatorname{argmax}}\ \log\det(\Theta) - \operatorname{trace}(\Theta S) - \lambda_n ||\Theta||_{1,\text{off}},$$

où $||\Theta||_{1,\text{off}} = \sum_{\substack{j,k=1 \\ j\neq k}}^{p} |\Theta_{jk}|$, S est la matrice de covariance empirique des observations du vecteur gaussien X et $\lambda_n > 0$ est le paramètre de la pénalisation Lasso. Ce problème d'optimisation est

convexe et possède une unique solution (voir Ravikumar *et al.* (2011) [83]). La particularité du graphical Lasso est de résoudre ce problème d'optimisation par la méthode de *pathwise descent coordinate*, ce qui le rend extrêmement rapide et performant (voir [44] pour plus de détails).

Dans notre cas, on ne peut pas obtenir la matrice de covariance empirique car les observations du vecteur X sont inaccessibles. Au lieu de calculer la matrice de covariance empirique des X comme la matrice de covariance empirique des observations $\mathbf{Y}$ du vecteur Y, on remplace cette matrice de covariance empirique S par l'estimateur $\widetilde{\Sigma}^{(n)}$ de Σ^* obtenu à l'étape 1.

Définition 6.1.2 (Estimateur $\widetilde{\Theta}^{(n)}$ de Θ^*)**.** *L'estimateur $\widetilde{\Theta}^{(n)}$ de la matrice de précision Θ^* est défini comme l'unique solution du problème d'optimisation convexe suivant :*

$$\widetilde{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\widetilde{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\mathit{off}}. \tag{6.3}$$

Les résultats théoriques de la section 6.2 concernent les propriétés des estimateurs $\widetilde{\Sigma}^{(n)}$ et $\widetilde{\Theta}^{(n)}$ respectivement définis en (6.2) et en (6.3) lorsque les points de troncature sont connus. En pratique, on ne connaît pas forcément ces points de troncature et on propose une étape préalable qui consiste à estimer ces points de troncature.

Étape 0 : estimation des points de troncature

Dans la littérature présentée en introduction de ce chapitre, les points de troncature sont supposés connus et leur estimation n'est donc pas étudiée. Cependant, en pratique, il est fréquent que ces points de troncature soient inconnus. Pour tout $j \in \{1, \dots, p\}$, on va estimer a_j et b_j par $\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ définis par :

$$\left.\begin{aligned} \hat{a}_j^{(n)} &:= \min_{i=1,\dots,n} \{Y_j^{(i)} : Y_j^{(i)} \neq 0\} \\ \hat{b}_j^{(n)} &:= \max_{i=1,\dots,n} \{Y_j^{(i)} : Y_j^{(i)} \neq 0\}. \end{aligned}\right\} \tag{6.4}$$

Si l'ensemble $\{Y_j^{(i)} : Y_j^{(i)} \neq 0\}$ est vide, on pose $\hat{a}_j^{(n)} = \hat{b}_j^{(n)} := 0$.

L'étape 1 utilise alors ces estimateurs et la matrice de covariance est estimée terme à terme comme suit :

$$\begin{aligned} \widehat{\Sigma}_{jk}^{(n)} &= \underset{|\sigma| \leq 1}{\operatorname{argmax}} \ L_{jk}^{(n)}(\sigma, \mathbf{y}, \hat{a}^{(n)}, \hat{b}^{(n)}) \\ &= \underset{|\sigma| \leq 1}{\operatorname{argmax}} \ \frac{1}{n} L_{jk}^{(n)}(\sigma, \mathbf{y}, \hat{a}^{(n)}, \hat{b}^{(n)}). \end{aligned} \tag{6.5}$$

L'estimateur de Θ^* est alors l'unique solution du problème d'optimisation :

$$\widehat{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\widehat{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\mathrm{off}}. \tag{6.6}$$

La section suivante comporte des résultats de convergence asymptotique des estimateurs $\widetilde{\Sigma}^{(n)}$ et $\widetilde{\Theta}^{(n)}$ ainsi que quelques résultats de convergence des estimateurs $\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ des points de troncature définis en (6.4). Nous ne présentons en revanche pas de résultats théoriques concernant les estimateurs $\widehat{\Sigma}^{(n)}$ et $\widehat{\Theta}^{(n)}$ obtenus après estimation des points de troncature. Ces estimateurs seront brièvement étudiés empiriquement dans la section 6.3.

6.2 Résultats de convergence

Dans cette section, on va présenter des résultats de convergence concernant différents estimateurs : les estimateurs $\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ des points de troncature, et les estimateurs $\widetilde{\Sigma}^{(n)}$ et $\widetilde{\Theta}^{(n)}$ de la matrice de covariance et de la matrice de précision respectivement (quand les points de troncature sont connus). L'objectif final est d'étudier les propriétés de l'estimateur $\widetilde{\Theta}^{(n)}$, obtenu par le problème d'optimisation décrit en (6.3), notamment celles concernant la récupération de la structure théorique de graphe, c'est-à-dire des conditions sous lesquelles l'estimateur $\widetilde{\Theta}^{(n)}$ retrouve correctement la nullité/non-nullité des coefficients de Θ^*. Ravikumar *et al.* (2001) [83] étudient le problème d'optimisation lié à la procédure du graphical Lasso et proposent un résultat qui requiert des conditions sur la vitesse de convergence de l'estimateur de la matrice de covariance. Notre estimateur est obtenu par maximisation de la log-vraisemblance des couples de variables du vecteur tronqué Y. Beaucoup de résultats théoriques ont été proposés pour les M-estimateurs, à commencer par Fisher (1922) [43] et (1925) [34] concernant l'efficacité asymptotique de ceux-ci. Par la suite, les résultats obtenus s'appuient surtout sur l'hypothèse de convexité (ou forte convexité) de la fonction cible (Candès et Tao (2005) [15], Candès et Tao (2007) [14], Negahban *et al.* (2012) [75]). Ces résultats ne s'appliquent toutefois pas dans notre cas, la log-vraisemblance n'étant pas concave.

Dans un premier temps, on présente quelques résultats asymptotiques sur les estimateurs des points de troncature. Puis on donne un résultat théorique sur la vitesse de convergence de $\widetilde{\Sigma}^{(n)}$, à l'aide de résultats de Mei *et al.* (2017) [68], valables dans un cadre non-convexe. Enfin, on expose les résultats théoriques concernant la vitesse de convergence de $\widetilde{\Theta}^{(n)}$ ainsi que sur sa parcimonie quant à l'estimation de la structure de graphe.

6.2.1 Estimateurs des points de troncature

Dans cette partie, on s'intéresse aux propriétés des estimateurs $(\hat{a}_j^{(n)})_{j\in[\![1,p]\!]}$ et $(\hat{b}_j^{(n)})_{j\in[\![1,p]\!]}$ des points de troncature définis en (6.4).

Proposition 6.2.1. *Soit $j \in \{1,\dots,p\}$.*

1. *$\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ sont fortement consistants :*

$$\hat{a}_j^{(n)} \xrightarrow[n\to+\infty]{p.s.} a_j,$$
$$\hat{b}_j^{(n)} \xrightarrow[n\to+\infty]{p.s.} b_j.$$

2. *On a les convergences en loi suivantes :*

$$n(a_j - \hat{a}_j^{(n)}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{E}\left(\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{a_j^2}{2}\Big)\right),$$
$$n(b_j - \hat{b}_j^{(n)}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{E}\left(\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{b_j^2}{2}\Big)\right),$$

où $\mathcal{E}(\lambda)$ désigne la loi exponentielle de paramètre $\lambda > 0$.

Démonstration. Soit $j \in \{1,\dots,p\}$. On montre les résultats pour $\hat{b}_j^{(n)}$.

1. Soit $0 < \epsilon < b_j - a_j$ et montrons que $\mathbb{P}\Big(\limsup_n \{|\hat{b}_j^{(n)} - b_j| > \epsilon\}\Big) = 0$. On étudie pour cela la série $\sum_{n\geq 1} \mathbb{P}\Big(|\hat{b}_j^{(n)} - b_j| > \epsilon\Big)$. Posons $N_{n,j} := \text{card}\{Y_j^{(i)} : Y_j^{(i)} \neq 0\}$ et $\Phi_j := \Phi(b_j) - \Phi(a_j)$ avec Φ la fonction de répartition de $\mathcal{N}(0,1)$. Alors, en remarquant que $\hat{b}_j^{(n)} \leq b$, on a :

$$\begin{aligned}
\mathbb{P}(|\hat{b}_j^{(n)} - b_j| > \epsilon) &= \mathbb{P}(\hat{b}_j^{(n)} < b_j - \epsilon) \\
&= \sum_{k=0}^{n} \mathbb{P}(N_{n,j} = k)\mathbb{P}(\hat{b}_j^{(n)} < b_j - \epsilon | N_{n,j} = k) \\
&= \sum_{k=1}^{n} \binom{n}{k} \Phi_j^k (1-\Phi_j)^{n-k} \left(\frac{\Phi(b_j - \epsilon) - \Phi(a_j)}{\Phi_j}\right)^k + (1-\Phi_j)^n \mathbb{1}_{b_j - \epsilon > 0} \\
&= \sum_{k=0}^{n} \binom{n}{k} (1-\Phi_j)^{n-k} (\Phi(b_j - \epsilon) - \Phi(a_j))^k - (1-\Phi_j)^n + (1-\Phi_j)^n \mathbb{1}_{b_j - \epsilon > 0} \\
&= \Big(1 - (\Phi(b_j) - \Phi(b_j - \epsilon))\Big)^n - (1-\Phi_j)^n \mathbb{1}_{b_j - \epsilon \leq 0}. \qquad (6.7)
\end{aligned}$$

Peu importe le signe de $b_j - \epsilon$, $\mathbb{P}(|\hat{b}_j^{(n)} - b_j| > \epsilon)$ est dans les deux cas le terme d'une série convergente. Ainsi, le lemme de Borel-Cantelli permet de conclure que $\mathbb{P}\Big(\limsup_n \{|\hat{b}_j^{(n)} - b_j| > \epsilon\}\Big) = 0$ et $\hat{b}_j^{(n)}$ converge presque sûrement vers b_j.

2. Soit $t \in \mathbb{R}$ et $n \in \mathbb{N}^*$ tel que $0 < \dfrac{t}{n} \leq b_j - a_j$.

$$\begin{aligned}
\mathbb{P}\Big(n(b_j - \hat{b}_j^{(n)}) \leq t\Big) &= \mathbb{P}\Big(\hat{b}_j^{(n)} \geq b_j - \frac{t}{n}\Big) \\
&= \Big(1 - \mathbb{P}\Big(\hat{b}_j^{(n)} < b_j - \frac{t}{n}\Big)\Big) \\
&= 1 - \left(1 - \Big(\Phi(b_j) - \Phi(b_j - \frac{t}{n})\Big)\right)^n + \Big(1 - \Phi(b_j) + \Phi(a_j)\Big)^n \mathbb{1}_{b_j - \frac{t}{n} \leq 0} \\
&\qquad \text{d'après (6.7)} \\
&= 1 - \exp\left(n \log\Big(1 - \Phi(b_j) + \Phi(b_j - \frac{t}{n})\Big)\right) \\
&\quad + \exp\left(n \log\Big(1 - \Phi(b_j) + \Phi(a_j)\Big)\right) \mathbb{1}_{b_j - \frac{t}{n} \leq 0} \\
&\underset{n \to +\infty}{\sim} 1 - \exp(t\Phi'(b_j)),
\end{aligned}$$

car $\Phi(b_j) - \Phi(b_j - \dfrac{t}{n}) = \dfrac{t}{n}\Phi'(b_j) + o(\dfrac{t}{n})$.

□

Corollaire 6.2.1. *Soit $j \in \{1, \ldots, p\}$ et $k \in \mathbb{N}$. Les moments d'ordre k de $\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ vérifient :*

$$\mathbb{E}\Big((\hat{a}_j^{(n)})^k\Big) \underset{n \to +\infty}{\longrightarrow} a_j^k,$$
$$\mathbb{E}\Big((\hat{b}_j^{(n)})^k\Big) \underset{n \to +\infty}{\longrightarrow} b_j^k.$$

En particulier, $\hat{a}_j^{(n)}$ *et* $\hat{b}_j^{(n)}$ *sont asymptotiquement sans biais et les variances de* $\hat{a}_j^{(n)}$ *et* $\hat{b}_j^{(n)}$ *convergent vers 0.*

Démonstration. Soit $j \in \{1, \dots, p\}$ et $k \in \mathbb{N}$. On a :

— $\hat{a}_j^{(n)} \xrightarrow[n\to+\infty]{\text{p.s.}} a_j$ donc $(\hat{a}_j^{(n)})^k \xrightarrow[n\to+\infty]{\text{p.s.}} a_j^k$

— Pour tout $n \in \mathbb{N}^*$, $|(\hat{a}_j^{(n)})^k| \leq \max(|a_j|, |b_j|)^k$.

Le théorème de convergence dominée permet de conclure. □

6.2.2 Estimateur de la matrice de covariance

On suppose ici que les points de troncature a_j et b_j sont connus pour tout $j \in \{1, \dots, p\}$ et l'estimateur $\widetilde{\Sigma}^{(n)}$ de la matrice de covariance Σ^* est alors donné par (6.2). On énonce au préalable deux hypothèses :

(H1) Pour tout $j < k$, $|\Sigma^*_{jk}| \neq 1$. Ainsi, il existe $\delta > 0$ tel que pour tout $j < k$, $|\Sigma^*_{jk}| < 1 - \delta$.

(H2) Soit $j < k$. On considère l'application $g : \sigma \in [-1+\delta, 1-\delta] \mapsto \mathbb{E}_{\Sigma^*_{jk}}\left(L_{jk}^{(n)}(\sigma, \mathbf{Y})\right)$. Alors, on suppose que :

- $-1 + \delta$ et $1 - \delta$ ne sont pas des points critiques de g,
- g admet un nombre fini de points critiques,
- tous les points critiques de g, différents de Σ^*_{jk}, sont non-dégénérés, c'est-à-dire :

$$\text{pour tout } \sigma \neq \Sigma^*_{jk},\ g'(\sigma) = 0 \Rightarrow g''(\sigma) \neq 0.$$

Notons que Σ^*_{jk} est un point critique non-dégénéré de g. Ceci sera démontré dans la preuve de la proposition 6.2.2 (voir équations (6.14)).

Voici un premier résultat concernant la vitesse de convergence de l'estimateur $\widetilde{\Sigma}^{(n)}$ de la matrice de covariance :

Proposition 6.2.2. *On suppose les hypothèses* ***(H1)*** *et* ***(H2)*** *énoncées précédemment. Soit* $0 < \rho < 1$. *Il existe des constantes* B, C *et* D *connues (qui dépendent de* δ, $(a_j)_{j\in[\![1,p]\!]}$, $(b_j)_{j\in[\![1,p]\!]}$*) telles que si* n *satisfait* $\dfrac{n}{\log n} \geq C \log\left(\dfrac{B}{\rho}\right)$, *alors la matrice de covariance estimée* $\widetilde{\Sigma}^{(n)}$ *définie par* (6.2) *vérifie :*

$$\mathbb{P}\left(||\widetilde{\Sigma}^{(n)} - \Sigma^*||_\infty \geq D\sqrt{\frac{\log n}{n}\log\left(\frac{B}{\rho}\right)}\right) \leq \rho\frac{p(p-1)}{2},$$

où $||A||_\infty = \max\limits_{j,k\in\{1,\dots,p\}} |A_{jk}|$ *est la norme infinie de la matrice* A *vue comme un élément de* $\mathbb{R}^{p^2}$.

La preuve de la proposition 6.2.2 utilise des résultats (principalement le théorème 2) de Mei *et al.* (2017) [68]. Dans cet article, Mei et ses coauteurs étudient les propriétés des points critiques d'une fonction de coût, appelée aussi risque empirique (par exemple, une vraisemblance ou log-vraisemblance), dans un cas non-convexe.

Démonstration. (Proposition 6.2.2) Pour prouver cette proposition, on va d'abord énoncer trois lemmes dont les démonstrations sont reportées en annexe :

Lemme 6.2.1. *Il existe $\gamma > 0$ tel que pour tout $j < k$, si $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ et $\sigma \in [-1+\delta, 1-\delta]$, alors pour tout $a, b \in \{0,1\}$, $\phi_{ab,jk}(\sigma, y_j, y_k) \geq \dfrac{1}{\gamma}$.*

Lemme 6.2.2. *Il existe L_1, L_2 et $L_3 > 0$ tels que pour tout $j < k$, si $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ et $\sigma \in [-1+\delta, 1-\delta]$, alors pour tout $a, b \in \{0,1\}$,*

$$\left|\partial_\sigma^m \phi_{ab,jk}(\sigma, y_j, y_k)\right| \leq L_m, \text{ pour } m \in \{1,2,3\}.$$

Lemme 6.2.3. *Soit $j < k$.*

1. *Pour tout $\sigma \in [-1+\delta, 1-\delta]$ et pour tout $l \in \mathbb{N}^*$,*

$$\int_{\mathbb{R}^2} \partial_\sigma^l \mathcal{L}_{jk}(\sigma, y) \mathrm{d}\mu(y) = \partial_\sigma^l \int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y) \mathrm{d}\mu(y) = 0,$$

où μ est la mesure sur $\mathbb{R}^2$ définie par :

$$\mu := \delta_0 \otimes \delta_0 + \delta_0 \otimes \lambda + \lambda \otimes \delta_0 + \lambda \otimes \lambda, \tag{6.8}$$

où δ_a désigne la masse de Dirac en $a \in \mathbb{R}$ et λ la mesure de Lebesgue sur $\mathbb{R}$.

2. *Pour tout $\sigma \in [-1+\delta, 1-\delta]$ et pour tout $l \in \mathbb{N}^*$,*

$$\partial_\sigma^l \mathbb{E}_{\Sigma_{jk}^*}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\partial_\sigma^l \log \mathcal{L}_{jk}(\sigma, Y)\Big),$$

c'est-à-dire,

$$\partial_\sigma^l \int_{\mathbb{R}^2} \log \mathcal{L}_{jk}(\sigma, y) \mathcal{L}_{jk}(\Sigma_{jk}^*, y) \mathrm{d}\mu(y) = \int_{\mathbb{R}^2} \partial_\sigma^l \Big(\log \mathcal{L}_{jk}(\sigma, y)\Big) \mathcal{L}_{jk}(\Sigma_{jk}^*, y) \mathrm{d}\mu(y).$$

Soient maintenant $j < k$ fixés. En suivant les notations de [68], on pose :

$$\begin{aligned}
\ell_{jk}(\sigma, \mathbf{y}) &= \log \mathcal{L}_{jk}(\sigma, \mathbf{y}) \\
&= \sum_{a=0}^{1} \sum_{b=0}^{1} n_{ab}(y_j, y_k) \log\Big(\phi_{ab,jk}(\sigma, y_j, y_k)\Big) && (6.9) \\
\hat{R}_n(\sigma, \mathbf{Y}) &= \frac{1}{n} L_{jk}^{(n)}(\sigma, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\sigma, Y^{(i)}) && (6.10) \\
R(\sigma) &= \mathbb{E}_{\Sigma_{jk}^*}\Big(\hat{R}_n(\sigma, \mathbf{Y})\Big) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma, Y)\Big) && (6.11) \\
&= \frac{1}{n} g(\sigma).
\end{aligned}$$

Remarques :

— Pour alléger les notations, on s'affranchit des indices j et k et on note ℓ, $\hat{R}_n$ et R au lieu de ℓ_{jk}, $\hat{R}_{n,jk}$ et R_{jk}.

— $\partial_\sigma \ell(\sigma, Y)$ correspond en fait au score du modèle.

— L'équation (6.2) se réécrit comme :

$$\widetilde{\Sigma}_{jk}^{(n)} = \underset{|\sigma| \leq 1}{\operatorname{argmax}}\ \hat{R}_n(\sigma, \mathbf{y}).$$

— Le point 2 du lemme 6.2.3 se réécrit comme :
Pour tout $\sigma \in [-1+\delta, 1-\delta]$ et pour $l \in \{1,2\}$:

$$R^{(l)}(\sigma) = \partial_\sigma^l \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma, Y))\Big) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\partial_\sigma^l \ell(\sigma, Y))\Big). \tag{6.12}$$

Vérifions à présent les quatre hypothèses requises pour l'application du théorème 2 de [68].

(i) Gradient statistical noise. *La dérivée de ℓ par rapport à σ est τ^2-sous-gaussienne pour un certain $\tau > 0$, i.e. :*

$$\exists \tau > 0, \forall \sigma \in [-1+\delta, 1-\delta],\ \forall \lambda \in \mathbb{R},\ \mathbb{E}\left[\exp\left(\lambda\Big(\partial_\sigma \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma \ell(\sigma, Y))\Big)\right)\right] \leq \exp\Big(\frac{\tau^2\lambda^2}{2}\Big).$$

En effet, pour tout $y \in \prod_{j=1}^{p} [a_j, b_j]$ et $\sigma \in [-1+\delta, 1-\delta]$,

$$\begin{aligned}
\partial_\sigma \ell(\sigma, y) &= \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k) \frac{\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)} \\
\text{Donc : } \Big|\partial_\sigma \ell(\sigma, y)\Big| &\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k) \frac{\big|\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)\big|}{\big|\phi_{ab,jk}(\sigma, y_j, y_k)\big|} \\
&\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\gamma L_1 = \gamma L_1 \text{ d'après les lemmes 6.2.1 et 6.2.2.}
\end{aligned}$$

Ainsi, $\partial_\sigma \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma \ell(\sigma, Y))$ est centrée et bornée par $2\gamma L_1$. D'après le lemme A.0.1, elle est donc τ^2-sous-gaussienne pour $\tau = 2\gamma L_1$. L'hypothèse "Gradient statistical noise" est donc vérifiée.

(ii) Hessian statistical noise. *La dérivée seconde de ℓ par rapport à σ est τ^2-sous-exponentielle ($\tau = 2\gamma L_1$) :*

$$||\partial_\sigma^2 \ell(\sigma, Y)||_{\psi_1} \leq \tau^2,$$

où $||.||_{\psi_1}$ est la norme ψ_1 de Orlicz définie par $||X||_{\psi_1} := \sup_{k \geq 1} \frac{1}{k}\mathbb{E}\Big(\big|X - \mathbb{E}(X)\big|^k\Big)^{\frac{1}{k}}$.

Il existe plusieurs définitions équivalentes (à certaines constantes universelles près) de la sous-exponentialité. Ici, nous avons donné celle énoncée dans la définition 3 de l'appendice *A Some useful tools* de [68]. Plus de détails sont donnés dans les références relatives mentionnées dans cet appendice telles que Vershynin (2012) [98] ou Boucheron (2013) *et al.* [10].

Pour tout $y \in \prod_{j=1}^{p} [a_j, b_j]$ et $\sigma \in [-1+\delta, 1-\delta]$,

$$\begin{aligned}
\partial_\sigma^2 \ell(\sigma, y) &= \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{\partial_\sigma^2 \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)} - \Big(\frac{\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)}\Big)^2\Big) \\
\text{Donc : } \Big|\partial_\sigma^2 \ell(\sigma, y)\Big| &\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{|\partial_\sigma^2 \phi_{ab,jk}(\sigma, y_j, y_k)|}{|\phi_{ab,jk}(\sigma, y_j, y_k)|} + \Big(\frac{|\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)|}{|\phi_{ab,jk}(\sigma, y_j, y_k)|}\Big)^2\Big) \\
&\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)(\gamma L_2 + \gamma^2 L_1^2) \text{ d'après les lemmes 6.2.1 et 6.2.2} \\
&= \gamma L_2 + \gamma^2 L_1^2. \qquad (6.13)
\end{aligned}$$

Ainsi, $\partial_\sigma^2 \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma^2 \ell(\sigma, Y))$ est bornée par $2(\gamma L_2 + \gamma^2 L_1^2)$. Donc pour tout $k \geq 1$,

$$\frac{1}{k}\mathbb{E}\Big(\big|\partial_\sigma^2 \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma^2 \ell(\sigma, Y))\big|^k\Big)^{\frac{1}{k}} \leq \frac{2}{k}(\gamma L_2 + \gamma^2 L_1^2),$$

et ainsi, $||\partial_\sigma^2 \ell(\sigma, Y)||_{\psi_1} \leq 2(\gamma L_2 + \gamma^2 L_1^2) \leq \tau^2 = 4\gamma^2 L_1^2$ quitte à choisir L_1 et/ou γ plus grands. Donc $\partial_\sigma^2 \ell(\sigma, Y)$ est bien τ^2-sous-exponentielle. L'hypothèse "Hessian statistical noise" est donc vérifiée.

(iii) Hessian regularity.

1. *La dérivée seconde de R (défini en (6.11)) est bornée en un point :*

$$\textit{il existe } |\sigma^*| \leq 1-\delta \textit{ et } H > 0 \textit{ tels que } \Big|R''(\sigma^*)\Big| \leq H.$$

2. *La dérivée seconde de ℓ par rapport à σ est Lipschitz-continue avec une constante Lipschitz-intégrable par rapport à y, c'est-à-dire :*

$$\textit{il existe } J^* > 0 \textit{ tel que } \mathbb{E}[J(Y)] \leq J^*,$$

où $J(y) = \sup_{\substack{|\sigma_1|,|\sigma_2| \leq 1-\delta \\ \sigma_1 \neq \sigma_2}} \frac{|\partial_\sigma^2 \ell(\sigma_1, y) - \partial_\sigma^2 \ell(\sigma_2, y)|}{|\sigma_1 - \sigma_2|}$.

3. $H \leq \tau^2$ *et* $J^* \leq \tau^3$.

Tout d'abord, on a $R''(\sigma) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\partial_\sigma^2 \ell(\sigma, Y)\Big)$ d'après le point 2 du lemme 6.2.3 et (6.12). D'après (6.13), $\Big|\partial_\sigma^2 \ell(\sigma, Y)\Big| \leq \gamma L_2 + \gamma^2 L_1^2$ pour tout $\sigma \in [-1+\delta, 1-\delta]$, donc n'importe quel $|\sigma^*| \leq 1-\delta$ et $H = \gamma L_2 + \gamma^2 L_1^2$ conviennent. Par ailleurs, quitte à choisir L_1 et/ou γ plus grands, on a bien $H \leq \tau^2 = 4\gamma^2 L_1^2$.

Pour tout $y \in \prod_{j=1}^{p}[a_j, b_j]$ et $\sigma \in [-1+\delta, 1-\delta]$, on a, en allégeant les notations :

$$\partial_\sigma^3 \ell(\sigma, y) = \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{\partial_\sigma^3 \phi_{ab,jk}}{\phi_{ab,jk}} - 3\frac{\partial_\sigma \phi_{ab,jk}\partial_\sigma^2 \phi_{ab,jk}}{\phi_{ab,jk}^2} + 2\Big(\frac{\partial_\sigma \phi_{ab,jk}}{\phi_{ab,jk}}\Big)^3\Big)$$

$$\begin{aligned}\text{Donc : } \left|\partial_\sigma^3 \ell(\sigma, y)\right| &\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)(\gamma L_3 + 3\gamma^2 L_1 L_2 + 2\gamma^3 L_1^3) \text{ (lemmes 6.2.1 et 6.2.2)} \\ &= \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3.\end{aligned}$$

Ainsi, pour tout $y \in \prod_{j=1}^{p}[a_j, b_j]$, $J(y) \leq \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3$ d'après l'inégalité des accroissements finis. $J^* = \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3$ convient et quitte à choisir L_1 et/ou γ plus grands, on a bien $J^* \leq \tau^3 = 8\gamma^3 L_1^3$. L'hypothèse "Hessian regularity" est donc vérifiée.

(iv) Morse. *Il existe $\epsilon > 0$ et $\eta > 0$ tels que R est (ϵ, η) fortement Morse, c'est-à-dire :*

1. *$|R'(\sigma)| > \epsilon$ pour tout σ tel que $|\sigma| = 1 - \delta$ et,*
2. *pour tout σ tel que $|\sigma| < 1 - \delta$, alors :*

$$|R'(\sigma)| \leq \epsilon \Rightarrow |R''(\sigma)| \geq \eta.$$

En d'autres termes, R vérifie cette condition si $-1+\delta$ et $1-\delta$ ne sont pas des points critiques de R et si R admet un nombre fini de points critiques, qui sont de plus non dégénérés, c'est-à-dire :

$$R'(\sigma) = 0 \Rightarrow R''(\sigma) \neq 0.$$

L'hypothèse **(H2)** implique le point 1. et le point 2. pour $\sigma \neq \Sigma_{jk}^*$. Il reste à montrer que Σ_{jk}^* est un point critique non-dégénéré.

Montrons que Σ_{jk}^* est un maximum global de R donc en particulier, un point critique de R. En effet, pour tout σ tel que $|\sigma| < 1$:

$$\left.\begin{aligned} R(\sigma) \leq R(\Sigma_{jk}^*) &\iff \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma, Y)\Big) \leq \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\Sigma_{jk}^*, Y)\Big) \\ &\iff \mathbb{E}_{\Sigma_{jk}^*}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) \leq \mathbb{E}_{\Sigma_{jk}^*}\Big(\log \mathcal{L}_{jk}(\Sigma_{jk}^*, Y)\Big) \\ &\text{car } \ell(\sigma, y) = \log \mathcal{L}_{jk}(\sigma, y) \text{ (définis en (6.1) et (6.9))} \\ &\iff \mathbb{E}_{\Sigma_{jk}^*}\Big(\log \frac{\mathcal{L}_{jk}(\sigma, Y)}{\mathcal{L}_{jk}(\Sigma_{jk}^*, Y)}\Big) \leq 0. \end{aligned}\right\} \tag{6.14}$$

D'après l'inégalité de Jensen,

$$\begin{aligned}\mathbb{E}_{\Sigma_{jk}^*}\left(\log \frac{\mathcal{L}_{jk}(\sigma, Y)}{\mathcal{L}_{jk}(\Sigma_{jk}^*, Y)}\right) &\leq \log \mathbb{E}_{\Sigma_{jk}^*}\left(\frac{\mathcal{L}_{jk}(\sigma, Y)}{\mathcal{L}_{jk}(\Sigma_{jk}^*, Y)}\right) \\ &= \log \int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) = 0, \end{aligned} \tag{6.15}$$

car $y \mapsto \mathcal{L}_{jk}(\sigma, y)$ est une densité de probabilité (voir (A.2)) par rapport à la mesure μ sur $\mathbb{R}^2$ définie en (6.8).

L'équation (6.15) implique que (6.14) est vérifiée. Σ^*_{jk} est donc un maximum global de R et on a alors $R'(\Sigma^*_{jk}) = 0$. Montrons que $R''(\Sigma^*_{jk}) \neq 0$:

$$\begin{aligned} R''(\Sigma^*_{jk}) &= \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^2_\sigma \ell(\Sigma^*_{jk}, Y)\Big) \text{ d'après le point 2 du lemme 6.2.3} \\ &= \mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\partial^2_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)} - \left(\frac{\partial_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}\right)^2\right) \\ &= -\mathbb{E}_{\Sigma^*_{jk}}\left(\left(\frac{\partial_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}\right)^2\right), \end{aligned}$$

car $\mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\partial^2_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}\right) = \int_{\mathbb{R}^2} \partial^2_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, y)\mathrm{d}\mu(y) = 0$ d'après le point 1 du lemme 6.2.3. Par l'absurde, si on suppose que $R''(\Sigma^*_{jk}) = 0$, alors $\partial_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y) = 0$ presque sûrement ce qui est absurde au regard de la définition de $\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)$ donnée en (6.1). Donc il existe $\epsilon > 0$ et $\eta > 0$ tels que R est (ϵ, η) fortement Morse. L'hypothèse "Morse" est donc vérifiée.

Pour chaque couple (j, k) tel que $j < k$, on peut à présent appliquer le théorème 2 de [68] à l'estimation du coefficient $\widetilde{\Sigma}^{(n)}_{jk}$. On obtient alors le résultat suivant :
Soient $j < k$ fixés et $0 < \rho < 1$. Il existe une constante universelle C_0 telle que si n vérifie $\frac{n}{\log n} \geq 4C_0\Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big) \vee 1\Big]\Big(\frac{\tau^2}{\epsilon^2} \vee \frac{\tau^4}{\eta^2} \vee \frac{\tau^2 L^2}{\eta^4}\Big)$ où $\tau = 2\gamma L_1$ et $L = \sup_{\sigma:|\sigma|\leq 1-\delta} |R^{(3)}(\sigma)|$,

$$\mathbb{P}\left(|\widetilde{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}| \leq \frac{2\tau}{\eta}\sqrt{C_0 \frac{\log n}{n}\ \Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big) \vee 1\Big]}\right) \geq 1 - \rho.$$

Et donc, pour $0 < \rho < 1$ et n tel que $\frac{n}{\log n} \geq 4C_0\Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big) \vee 1\Big]\Big(\frac{\tau^2}{\epsilon^2} \vee \frac{\tau^4}{\eta^2} \vee \frac{\tau^2 L^2}{\eta^4}\Big)$, alors :

$$\mathbb{P}\left(\big|\big|\widetilde{\Sigma}^{(n)} - \Sigma^*\big|\big|_\infty \leq \frac{2\tau}{\eta}\sqrt{C_0 \frac{\log n}{n}\ \Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big) \vee 1\Big]}\right) \geq 1 - \rho\frac{p(p-1)}{2},$$

où $||A||_\infty = \max_{j,k\in\{1,\dots,p\}} |A_{jk}|$ est la norme infinie de la matrice A vue comme un élément de $\mathbb{R}^{p^2}$. Ceci achève la preuve de la Proposition 6.2.2.

□

6.2.3 Estimateur de la matrice de précision

Avant de donner un second résultat concernant l'estimateur de la matrice de précision Θ^*, énonçons une troisième hypothèse :

(H3) Il existe $\alpha \in]0, 1]$ tel que :

$$\max_{e\in S^c} \big|\big|\Gamma^*_{eS}(\Gamma^*_{SS})^{-1}\big|\big|_1 = \big|\big|\big|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\big|\big|\big|_\infty \leq 1 - \alpha,$$

où :

— $\Gamma^* = \Sigma^* \otimes \Sigma^*$ où $\otimes$ désigne le produit de Kronecker. On a : $\Gamma^*_{(j,k),(l,m)} = \mathrm{cov}(X_j X_k, X_l X_m)$ et donc $\Gamma^*_{SS} \in \mathcal{M}_{s+p,s+p}(\mathbb{R})$,

— si $M \in \mathcal{M}_{r,m}(\mathbb{R})$, $A \subset [\![1,r]\!]$ et $B \subset [\![1,m]\!]$, M_{AB} désigne la matrice extraite $(m_{ij})_{i\in A, j\in B}$,

— $S = S(\Theta^*) := E(\Theta^*) \cup \{(1,1),\dots,(p,p)\}$ où $\Theta^* = (\Sigma^*)^{-1}$ et $E(\Theta^*) = \{(j,k) \in \{1,\dots,p\}^2,\ j \neq k,\ \Theta^*_{jk} \neq 0\}$. On pose : $s := |E(\Theta^*)|$, d'où $|S(\Theta^*)| = |E(\Theta^*)| + p = s + p$,

— $S^c = S^c(\Theta^*) = [\![1,p]\!]^2 \setminus S(\Theta^*)$,

— $||u||_1 = \sum_{j=1}^d |u_j|$ pour tout $u \in \mathbb{R}^d$ est la norme 1 usuelle,

— $|||U|||_\infty = \max\limits_{j=1,\dots,d} \sum\limits_{k=1}^m |U_{jk}|$ pour tout $U \in \mathcal{M}_{d,m}(\mathbb{R})$.

Remarque : $E(\Theta^*)$ correspond à l'ensemble des paires d'indices (j,k) telles que X_j et X_k sont liées par une arête, c'est-à-dire conditionnellement dépendantes aux autres variables :

$$\begin{aligned}(j,k) \in E(\Theta^*) &\iff (k,j) \in E(\Theta^*)\\ &\iff \Theta^*_{jk} (= \Theta^*_{kj}) \neq 0\\ &\iff X_j \not\perp\!\!\!\perp X_k | (X_l)_{l\neq j,k}.\end{aligned}$$

L'intuition sous-jacente de l'hypothèse **(H3)** est de limiter l'influence des termes relatifs à des "non-arêtes", indexés par S^c, sur les termes relatifs à des arêtes, indexés par S. On pourra se référer à [83] pour plus de détails.

La proposition suivante établit la consistance en norme infinie de l'estimateur $\widetilde{\Theta}^{(n)}$ de la matrice de précision, obtenu par la procédure Lasso décrite précédemment (6.3). On introduit au préalable quelques notations :

— d est le degré maximal du graphe défini par :

$$d = \max_{j=1,\dots,p} \Big| \{k \in [\![1,p]\!] : \Theta^*_{jk} \neq 0\} \Big|. \tag{6.16}$$

— $E(\widetilde{\Theta}^{(n)})$, κ_{Σ^*} et κ_{Γ^*} sont définis par :

$$E(\widetilde{\Theta}^{(n)}) := \{(j,k) \in \{1,\dots,p\}^2,\ j \neq k,\ \widetilde{\Theta}^{(n)}_{jk} \neq 0\},$$

$$\kappa_{\Sigma^*} := |||\Sigma^*|||_\infty = \max_{j=1,\dots,p} \sum_{k=1}^p |\Sigma^*_{jk}|, \tag{6.17}$$

$$\kappa_{\Gamma^*} := \left|\left|\left| \left(\Gamma^*_{SS}\right)^{-1} \right|\right|\right|_\infty. \tag{6.18}$$

Proposition 6.2.3. *On suppose l'hypothèse **(H3)** et on suppose qu'il existe des constantes B, C, D strictement positives et $c > 2$ tel que pour n tel que $\dfrac{n}{\log n} \geq C \log(Bp^c)$,*

$$\mathbb{P}\left(||\widetilde{\Sigma}^{(n)} - \Sigma^*||_\infty \geq D\sqrt{\frac{\log n}{n} \log(Bp^c)} \right) \leq \frac{p(p-1)}{2p^c}, \tag{6.19}$$

où $||A||_\infty = \max\limits_{j,k\in\{1,\dots,p\}} |A_{jk}|$ désigne la norme infinie de la matrice A vue comme un élément de $\mathbb{R}^{p^2}$.

Soit n *vérifiant* $\dfrac{n}{\log n} > D^2 \log(Bp^c) \max\Big\{\dfrac{\sqrt{C}}{D}, 6(1+8\alpha^{-1})d\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa_{\Sigma^*}^3\kappa_{\Gamma^*}^2\}\Big\}^2$, $\widetilde{\Theta}^{(n)}$ *l'unique solution de* (6.3) *et* $\lambda_n = \dfrac{8D}{\alpha}\sqrt{\dfrac{\log n}{n}\log(Bp^c)}$ *le paramètre de pénalisation de l'équation Lasso* (6.3)*. Alors, avec probabilité au moins* $1 - \dfrac{1}{p^{c-2}}$*, on a :*

(a) L'estimateur $\widetilde{\Theta}^{(n)}$ *de* Θ^* *satisfait :*

$$||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log(Bp^c)}.$$

(b) $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ *et l'arête* (j,k) *est correctement retrouvée dès que :*

$$|\Theta^*_{jk}| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log(Bp^c)}.$$

La démonstration s'appuie sur les résultats du théorème 1 de l'article de Ravikumar *et al.* (2011) [83], dans lequel ils étudient le problème de l'estimation de la matrice de précision dans un cadre général incluant la procédure du graphical Lasso dans le cadre gaussien multivarié.

Démonstration. (Proposition 6.2.3) Vérifions au préalable les hypothèses d'application de ce théorème.

• **Incoherence assumption.** Cette hypothèse correspond exactement à l'hypothèse **(H3)**.

• **Control of sampling noise.** Une lecture attentive de [83] montre que l'hypothèse *tail condition* du théorème 1 n'est pas nécessaire. L'hypothèse nécessaire est en fait un résultat plus faible, donné dans le lemme 8 de [83], et énoncé ci-dessous.
Il existe $v^* > 0$ *tel que pour tout* $c > 2$ *et* n *tel que* $\bar{\beta}_f(n, p^c) \leq \dfrac{1}{v_*}$*, on a :*

$$\mathbb{P}\Big[||\widetilde{\Sigma}^{(n)} - \Sigma^*||_\infty \geq \bar{\beta}_f(n,p^c)\Big] \leq \frac{1}{p^{c-2}},$$

où $\bar{\beta}_f(n,r) := \text{argmax}\{\beta : f(n,\beta) \leq r\}$ *pour une certaine fonction* $f(n,\beta)$*.*

En posant $f(n,\beta) = \dfrac{1}{B}\exp\left(\dfrac{n}{\log n}\Big(\dfrac{\beta}{D}\Big)^2\right)$ et $v_* = \dfrac{\sqrt{C}}{D}$ et en remarquant que $\dfrac{p(p-1)}{2} \leq p^2$, l'hypothèse (6.19) donne ce résultat. En effet :

— $\bar{\beta}_f(n,r) = \text{argmax}\{\beta : f(n,\beta) \leq r\} = D\sqrt{\dfrac{\log n}{n}\log(Br)}$

— $\dfrac{n}{\log n} \geq C\log(Bp^c) \iff \bar{\beta}_f(n,p^c) \leq \dfrac{D}{\sqrt{C}}$

L'hypothèse "Control of sampling noise" est donc vérifiée.

Enfin, posons $\bar{n}_f(\beta,r) := \text{argmax}\{n : f(n,\beta) \leq r\}$. On a alors que la condition $n > \bar{n}_f(\beta,r)$ est équivalente à $\dfrac{n}{\log n} > \log(Br)\dfrac{D^2}{\beta^2}$ puisque $f(n,\beta) \leq r \iff \dfrac{n}{\log n} \leq \log(Br)\dfrac{D^2}{\beta^2}$.

Nous sommes maintenant en mesure d'appliquer le théorème 1 de [83]. On obtient que pour $\lambda_n = \frac{8}{\alpha}\bar{\beta}_f(n, p^c)$ le paramètre de pénalisation de l'équation Lasso (6.3) et

$$n > \bar{n}_f\left(\frac{1}{\max\left\{v_*, 6(1+8\alpha^{-1})d\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}\right\}}, p^c\right),$$

on a, avec probabilité au moins $1 - \frac{1}{p^{c-2}}$:

(a) L'estimateur $\widetilde{\Theta}^{(n)}$ de Θ^* satisfait :

$$||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq 2(1+8\alpha^{-1})\kappa_{\Gamma^*}\bar{\beta}_f(n, p^c)$$

(b) $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ et l'arête (j,k) est correctement retrouvée si $|\Theta^*_{jk}| > 2(1+8\alpha^{-1})\kappa_{\Gamma^*}\bar{\beta}_f(n, p^c)$, où $\widetilde{\Theta}^{(n)}$ est l'unique solution de (6.3). □

Finalement, les propositions 6.2.2 et 6.2.3 fournissent le résultat suivant, qui précise la vitesse de convergence de l'estimateur de la matrice de précision ainsi que des conditions sous lesquelles le graphe théorique est correctement retrouvé :

Théorème 6.2.1. *On suppose les hypothèses **(H1)**, **(H2)** et **(H3)** énoncées précédemment. Soit $c > 2$, $\widetilde{\Theta}^{(n)}$ l'unique solution de (6.3) et α, d, κ_{Σ^*} et κ_{Γ^*} définis respectivement à l'hypothèse **(H3)**, en (6.16), en (6.17) et en (6.18). Il existe des constantes B, C et D connues (qui dépendent notamment de δ, $(a_j)_{j\in[\![1,p]\!]}$, $(b_j)_{j\in[\![1,p]\!]}$) telles que pour n vérifiant $\frac{n}{\log n} > D^2\log(Bp^c)\max\left\{\frac{\sqrt{C}}{D}, 6(1+8\alpha^{-1})d\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}\right\}^2$ et $\lambda_n = \frac{8D}{\alpha}\sqrt{\frac{\log n}{n}\log(Bp^c)}$ le paramètre de pénalisation de l'équation Lasso (6.3), on a, avec probabilité au moins $1 - \frac{1}{p^{c-2}}$:*

(a) L'estimateur $\widetilde{\Theta}^{(n)}$ de Θ^* satisfait :

$$||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log(Bp^c)}.$$

(b) $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ et l'arête (j,k) est correctement retrouvée dès que :

$$|\Theta^*_{jk}| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log(Bp^c)}.$$

Le paramètre c du théorème 6.2.1 est en fait un paramètre à définir par l'utilisateur. Plus c est grand, plus la probabilité pour laquelle les deux résultats du théorème tiennent sera grande. En revanche, de grandes valeurs de ce paramètre c conduisent également à de plus fortes exigences sur la taille n de l'échantillon.

6.3 Simulations

6.3.1 Paramètres de simulations

Dans la quasi-totalité de cette section (sauf mention contraire), on utilisera les paramètres de simulations suivants. On simule $n = 500$ observations d'un $p = 100$-vecteur gaussien X centré

et réduit dont la structure de graphe est une chaîne, c'est-à-dire que $X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_{100}$. Le graphe comporte alors 99 arêtes. Les données sont générées à l'aide de la fonction `huge.generator`, option `graph = "band"` du package `huge`.

Plusieurs points de troncature sont présentés par la suite :

- seuils identiques : $a = -0.5$ et $b = 2$,
- seuils symétriques : $a = -1.5$ et $b = 1.5$,
- seuils croissants : $a =$ `seq(-1,0, length = p)`, $b =$ `seq(0.5,3, length = p)`,
- seuils décroissants : $a = -1$, $b =$ `seq(2,0.5, length = p)`.

On suit ensuite la procédure d'estimation décrite dans la sous-section 6.1.3. Dans un cas, on suppose que les points de troncature sont connus et on utilise les estimateurs $\widetilde{\Sigma}^{(n)}$ et $\widetilde{\Theta}^{(n)}$ des matrices de covariance et précision, respectivement définis à l'étape 1 en (6.2) et à l'étape 2 en (6.3). Dans l'autre, on estime d'abord les points de troncature grâce aux estimateurs $\hat{a}_j^{(n)}$ et $\hat{b}_j^{(n)}$ présentés à l'étape 0 en (6.4) et on utilise les estimateurs $\widehat{\Sigma}^{(n)}$ et $\widehat{\Theta}^{(n)}$ respectivement définis en (6.5) et (6.6).

Pour l'étape 2, qui consiste à estimer la matrice de précision à l'aide du graphical Lasso, on utilise la fonction `huge` du package éponyme, option `method = "glasso"`. Malheureusement, les résultats théoriques exposés à la section 6.2 ne permettent pas de choisir explicitement le paramètre de pénalisation (de la même façon qu'ils ne fournissent pas explicitement le nombre d'observations nécessaires pour que ces résultats tiennent). Le package `huge` propose en revanche plusieurs méthodes pour le choix du paramètre de pénalisation λ optimal : la méthode "ebic", la méthode "stars" et la méthode "ric". On a essayé et comparé chacune de ces méthodes pour le choix du paramètre. Ces résultats sont présentés dans la sous-section 6.3.2.

Au travers de ces simulations, on cherche à répondre à différentes problématiques. Pour ce faire, on va expliciter les trois différentes procédures qui seront utilisées par la suite, à savoir "notre procédure - points de troncature connus", "notre procédure - points de troncature estimés/inconnus" et "graphical Lasso directement sur données tronquées". Notre procédure, qui consiste à remplacer la matrice de covariance empirique du vecteur gaussien X dans le graphical Lasso par un autre estimateur, possède deux variantes. Dans le premier cas, on connaît les points de troncature et l'estimateur de Σ^* utilisé est $\widetilde{\Sigma}^{(n)}$. Dans le second cas, on ne connaît pas les points de troncature. On les estime alors et l'estimateur de Σ^* utilisé est $\widehat{\Sigma}^{(n)}$. La troisième procédure utilisée est le graphical Lasso utilisé directement sur les données tronquées Y, qui consiste à remplacer la matrice de covariance empirique du vecteur gaussien X par la matrice de covariance empirique du vecteur tronqué Y. Voici dans l'ordre les problématiques qui seront abordées dans les sous-sections suivantes :

— Le choix du paramètre de pénalisation en pratique, à l'aide des méthodes proposées par le package `huge`. La même méthode convient-elle pour chacune des trois procédures ? Quelle est la plus adaptée ?

— L'efficacité de notre procédure. Utiliser nos estimateurs pour la matrice de covariance améliore-t-il vraiment l'estimation du réseau ?

— L'impact de l'estimation des points de troncature sur l'estimation du réseau.

— L'impact des points de troncature, c'est-à-dire comment les valeurs des points de troncature impactent la détection des arêtes du réseau.

— Que donne notre procédure sur d'autres structures de graphes ?

Pour étudier les performances de ces procédures sur l'estimation du réseau, on répète la procédure étudiée 50 fois et on s'intéresse aux taux de détection de chacune des $\binom{100}{2} = 4950$ arêtes potentielles.

6.3.2 Choix du paramètre de pénalisation

Dans un premier temps, on va exposer les résultats concernant le choix du paramètre de pénalisation par ces trois méthodes. Les résultats étant similaires pour les différents points de troncature, on va se contenter d'exposer les résultats pour la configuration "seuils identiques" : $a = -0.5$ et $b = 2$. On présente les résultats de notre procédure, points de troncature connus et les résultats de la procédure graphical Lasso directement sur les données tronquées.

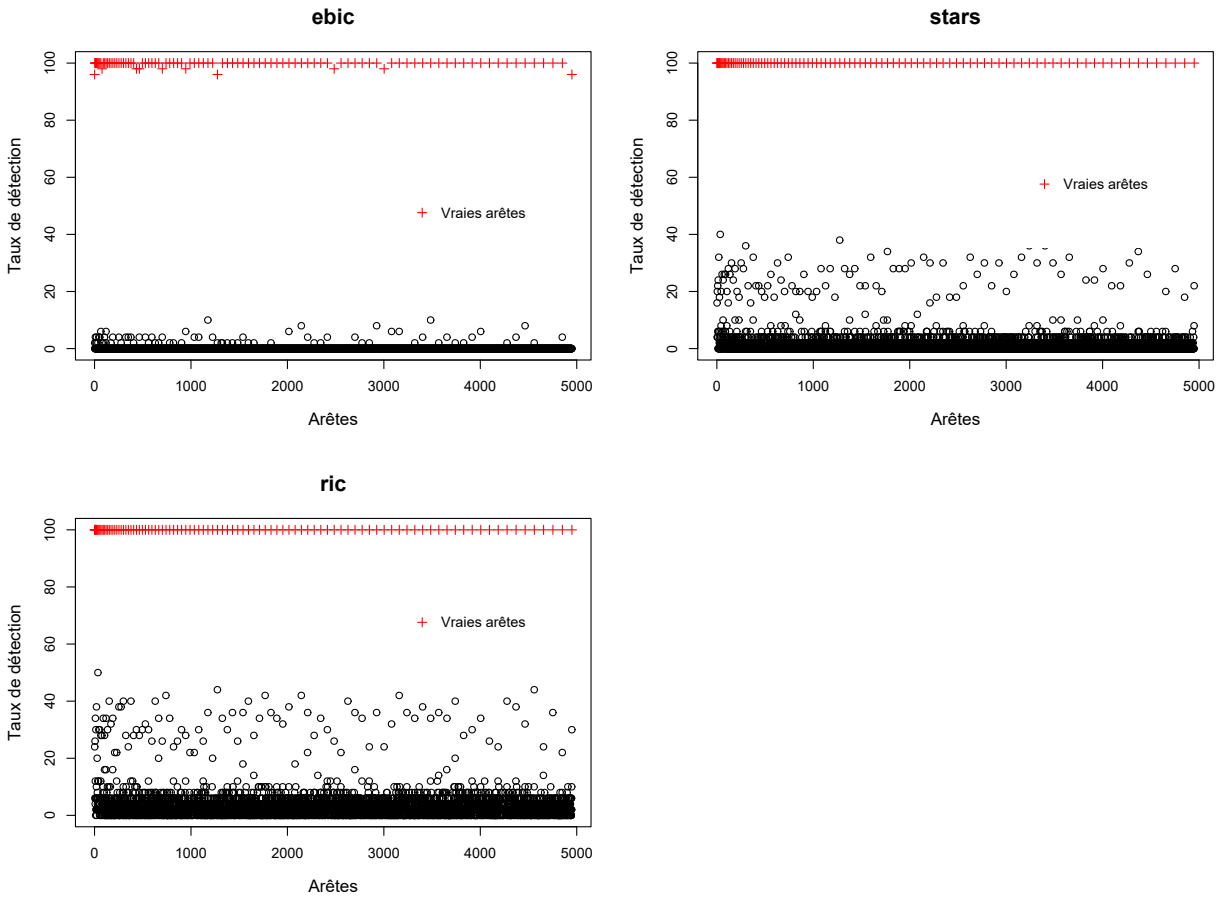


FIGURE 6.1 – Comparaison des taux de détection obtenus par les méthodes "ebic", "stars" et "ric" pour le choix du paramètre de pénalisation avec notre procédure, points de troncature connus. Les points de troncature sont connus et valent $a = -0.5$ et $b = 2$. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

Résultats et commentaires. La figure 6.1 représente les taux de détection de chaque arête avec notre procédure pour différentes méthodes de sélection du paramètre de pénalisation. Les

vraies arêtes sont représentées en rouge. On peut constater que la méthode “ebic” semble donner les résultats les plus satisfaisants : les vraies arêtes sont bien détectées et il y a peu de faux positifs. En comparaison, les méthodes “stars” et “ric” détectent beaucoup (à tort) les autres arêtes. Les résultats pour notre procédure, points de troncature estimés, ne sont pas présentés mais sont similaires. Pour la suite des résultats exposés, on utilisera la méthode “ebic” pour le choix du paramètre de pénalisation dans notre procédure, points de troncature connus ou non.

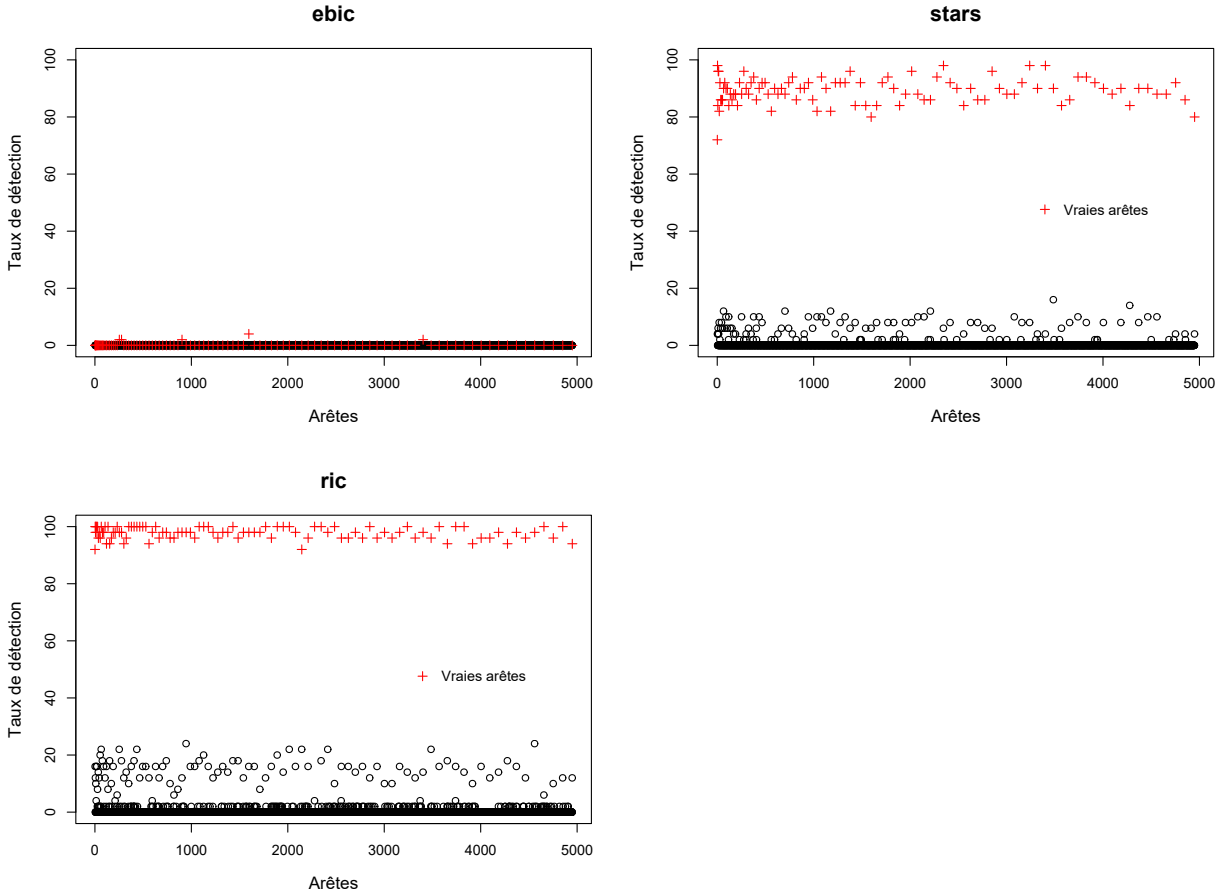
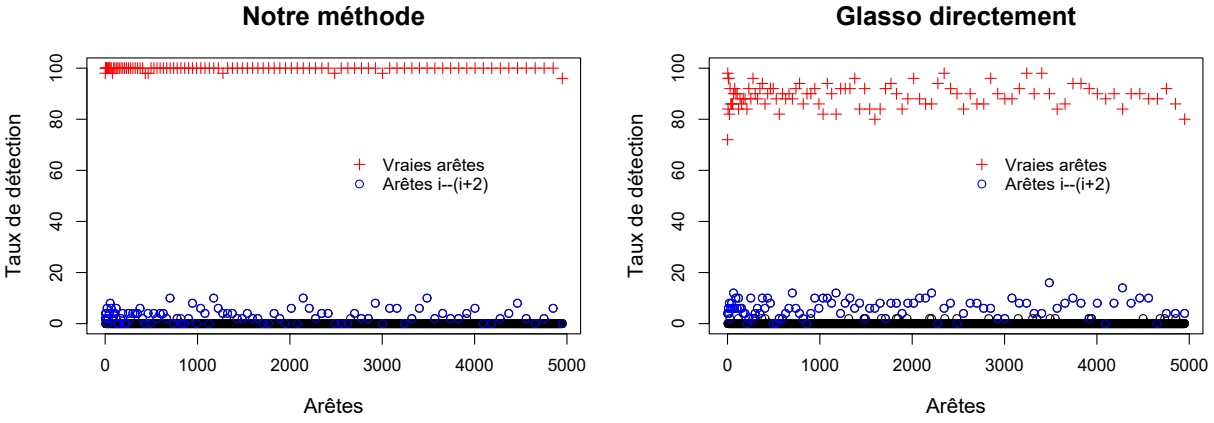


FIGURE 6.2 – Comparaison des taux de détection obtenus par les méthodes “ebic”, “stars” et “ric” pour le choix du paramètre de pénalisation par le graphical Lasso directement sur les données tronquées. Les points de troncature sont connus et valent $a = -0.5$ et $b = 2$. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

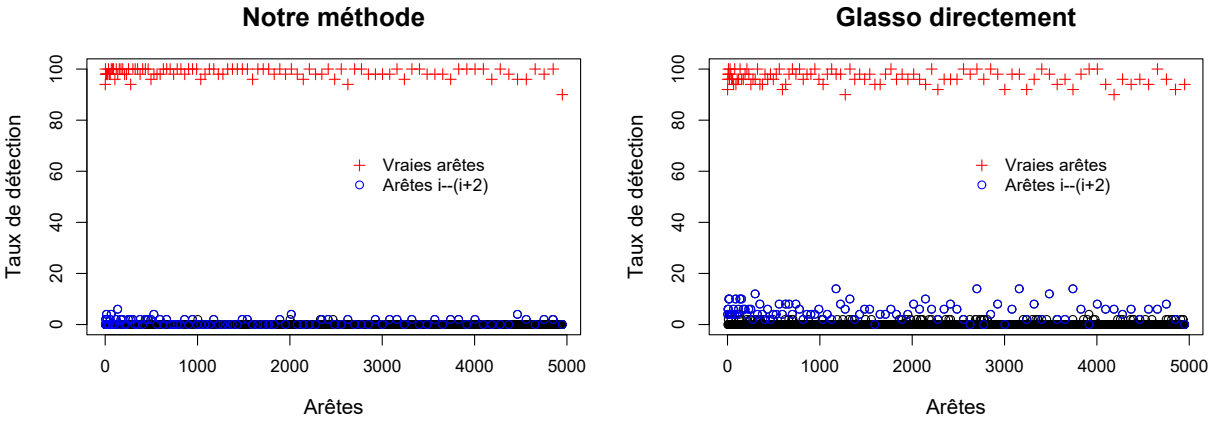
Résultats et commentaires. La figure 6.2 représente les taux de détection de chaque arête pour différentes méthodes de sélection du paramètre de pénalisation. Dans ce cas, on utilise la procédure du graphical Lasso directement sur les données tronquées. On peut constater que la méthode “ebic” est ici très décevante et ne détecte quasiment aucune arête. Les méthodes “stars” et “ric” donnent des résultats plus satisfaisants et comparables : les vraies arêtes sont bien détectées malgré un certain nombre de faux positifs. Pour la suite des résultats exposés, on utilisera la méthode “stars” pour le choix du paramètre de pénalisation pour le graphical Lasso directement sur les données tronquées.

6.3.3 Efficacité de la procédure

On s'intéresse ici à illustrer l'efficacité de notre procédure. Pour ce faire, on compare les taux de détection de chacune des potentielles arêtes obtenus par notre méthode à ceux obtenus par le graphical Lasso directement sur les données tronquées. Pour notre méthode, on présente les résultats dans le cas où les points de troncature sont estimés, qui diffèrent très peu des résultats dans le cas où les points de troncature sont connus en amont (voir sous-section 6.3.4). Le paramètre de pénalisation est choisi à l'aide la méthode "ebic" pour notre méthode et "stars" pour le graphical Lasso sur les données tronquées (voir sous-section 6.3.2).



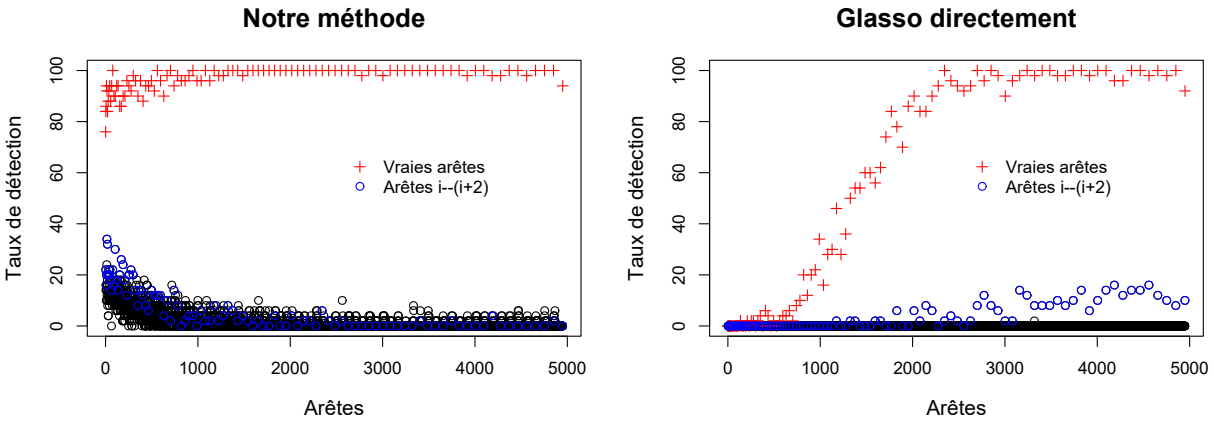
(a) Seuils identiques : $a = -0.5$ et $b = 2$.



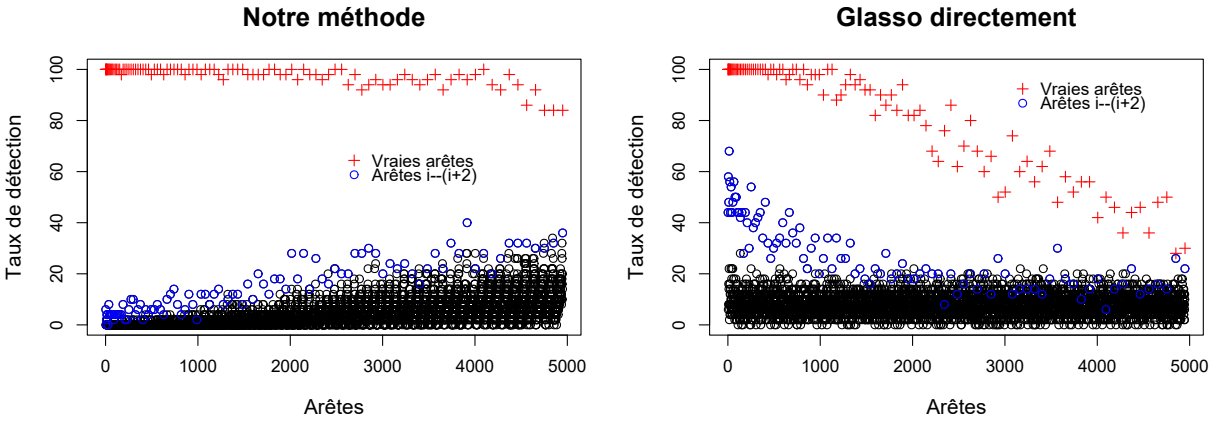
(b) Seuils symétriques : $a = -1.5$, $b = 1.5$.

FIGURE 6.3 – Comparaison des taux de détection obtenus par notre méthode (points de troncature estimés) et par le graphical Lasso appliqué directement sur les données tronquées. Les configurations des points de troncature sont ici "seuils identiques" et "seuils symétriques". Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables. En rouge sont représentées les 99 vraies arêtes, en bleu les arêtes du type $X_i - X_{i+2}$.

Résultats et commentaires. Les figures 6.3 et 6.4 illustrent les taux de détection pour notre



(a) Seuils croissants : $a =$ `seq(-1,0, length = p)`, $b =$ `seq(0.5,3, length = p)`.



(b) Seuils décroissants : $a = -1$, $b =$ `seq(2,0.5, length = p)`.

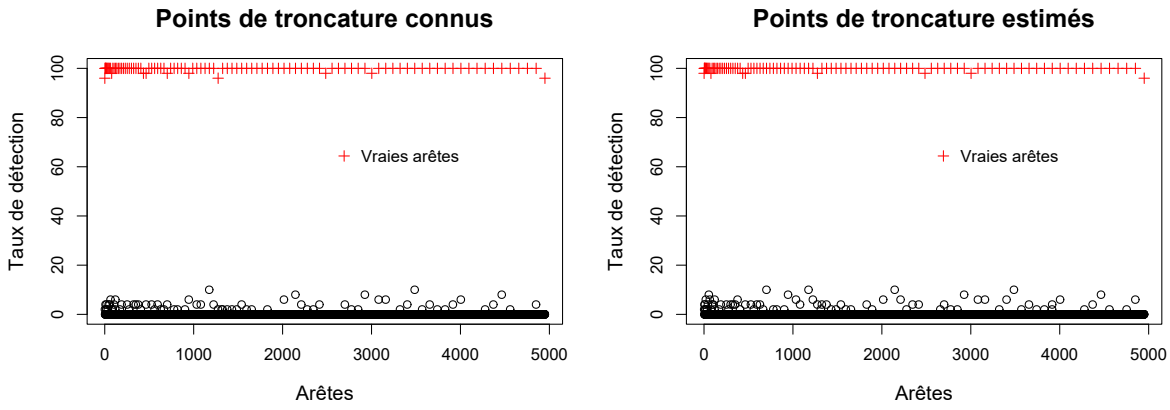
FIGURE 6.4 – Comparaison des taux de détection obtenus par notre méthode (points de troncature estimés) et par le graphical Lasso appliqué directement sur les données tronquées. Les configurations des points de troncature sont ici "seuils croissants" et "seuils décroissants". Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables. En rouge sont représentées les 99 vraies arêtes, en bleu les arêtes du type $X_i - X_{i+2}$.

méthode et pour le graphical Lasso (abrégé en Glasso) utilisé directement sur les données tronquées Y. On représente les taux de détection de chacune des potentielles arêtes en déroulant la matrice avec un `upper.tri`, c'est-à-dire que la première arête dont le taux de détection est représenté est l'arête $X_2 \longleftrightarrow X_1$, puis $X_3 \longleftrightarrow X_1$, $X_3 \longleftrightarrow X_2$, $X_4 \longleftrightarrow X_1$, ..., $X_4 \longleftrightarrow X_3$ etc. Les 99 vraies arêtes théoriques, c'est-à-dire les arêtes $X_i \longleftrightarrow X_{i+1}$, sont représentées en rouge et dans l'ordre $X_1 \longleftrightarrow X_2$, $X_2 \longleftrightarrow X_3$, ..., $X_{99} \longleftrightarrow X_{100}$. Les arêtes de la forme $X_i \longleftrightarrow X_{i+2}$, qui ne sont pas des vraies arêtes, sont représentées en bleu. Il s'agit d'interactions qui peuvent être relativement fortes (à cause du lien indirect via X_{i+1}), qui sont de ce fait intéressantes à étudier.
La figure 6.3 présente les configurations des points de troncature "seuils identiques" et "seuils symétriques". Dans les deux cas, on constate que notre méthode fournit de meilleurs résultats : les vraies arêtes sont mieux détectées et les autres arêtes le sont moins. Par exemple, dans la configuration "seuils identiques", les vraies arêtes sont toutes détectées plus de 96% alors que le Glasso les détecte plutôt entre 80% et 100% : 66 de ces arêtes sont détectées au plus à 90%. La différence de détection est moins nette dans la configuration "seuils symétriques", ceci est probablement dû au plus faible pourcentage de données nulles dans les données tronquées Y (environ 13% pour les "seuils symétriques" contre 33% pour les "seuils identiques"). Les arêtes de la forme $X_i \longleftrightarrow X_{i+2}$ ont tendance à être davantage détectées avec le Glasso.
La figure 6.4 présente les configurations des points de troncature "seuils croissants" et "seuils décroissants". L'efficacité de notre procédure est encore plus nette dans ces configurations. En effet, les vraies arêtes sont beaucoup mieux détectées par notre méthode (plus de 80% dans les deux configurations alors que respectivement 60 et 34 vraies arêtes ne sont détectées qu'au plus 80% par le graphical Lasso). Ces différences de détection s'expliquent en partie par le nombre de zéros dans les données tronquées Y : dans la configuration décroissante notamment, le taux de zéros croît de 20% à 50% selon les variables. Ainsi, les arêtes impliquant les variables dont le taux de zéros est fort (c'est-à-dire les arêtes $X_i \longleftrightarrow X_{i+1}$ pour les i proches de 100) ont tendance à être moins bien détectées, et ce phénomène est encore plus marqué avec le graphical Lasso directement sur les données inflatées en zéro par troncature. Par ailleurs, les autres arêtes (les "fausses") ont tendance à être légèrement moins détectées avec le graphical Lasso sauf en ce qui concerne les arêtes de la forme $X_i \longleftrightarrow X_{i+2}$. Pour notre méthode, les "fausses" arêtes sont davantage détectées quand les points de troncature induisent un fort taux de zéros pour les variables impliquées.
La sous-section 6.3.5 apporte une comparaison plus poussée des taux de détection selon les configurations des points de troncature.

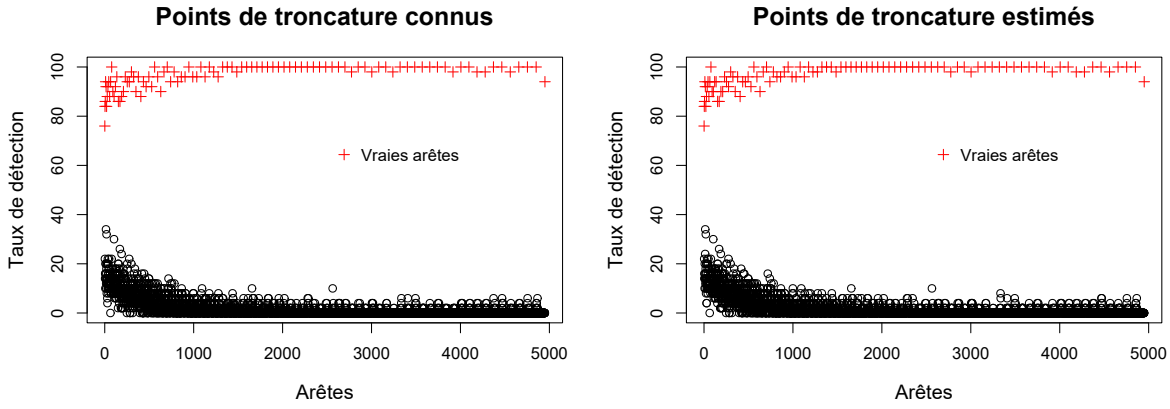
6.3.4 Impact de l'estimation des points de troncature

Dans cette partie, on va présenter des graphiques illustrant l'impact de l'estimation des points de troncature. Pour plusieurs configurations des points de troncature (seuils identiques, croissants et décroissants), on va comparer les taux de détection dans le cas où les points de troncature sont connus au cas où les points de troncature sont inconnus et estimés à l'aide de l'étape 0 présentée à la sous-section 6.1.3. Dans ces deux cas, le paramètre de pénalisation λ est choisi grâce à la méthode "ebic" (voir sous-section 6.3.2).

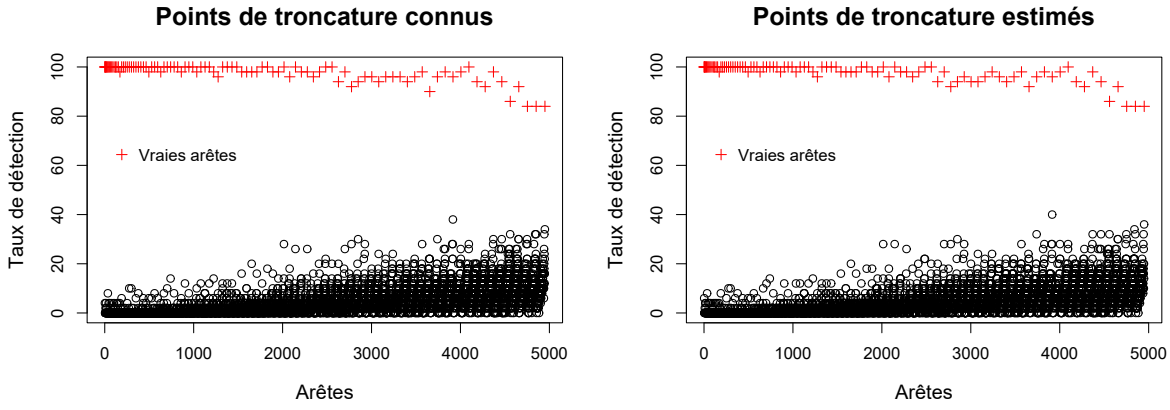
Résultats et commentaires. La figure 6.5 illustre l'impact de l'estimation des points de troncature (étape 0) dans notre procédure pour différentes configurations des points de troncature à savoir les configurations "seuils identiques", "seuils croissants" et "seuils décroissants". Pour chacune de ces configurations, on peut constater que l'impact de l'estimation des points de tron-



(a) Seuils identiques : $a = -0.5$ et $b = 2$.



(b) Seuils croissants : $a =$ `seq(-1,0, length = p)`, $b =$ `seq(0.5,3, length = p)`.



(c) Seuils décroissants : $a = -1$, $b =$ `seq(2,0.5, length = p)`.

FIGURE 6.5 – Comparaison des taux de détection lorsque les points de troncature sont connus et des taux de détection lorsque les points de troncature sont estimés. On s'intéresse aux configurations "seuils identiques", "seuils croissants" et "seuils décroissants" pour les points de troncature. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

cature est quasiment nul. En fait, on constate peu d'écarts entre les taux de détection des arêtes. Dans ces simulations, on dispose toutefois de $n = 500$ observations, ce qui permet de fournir de bonnes estimations des points de troncature et ce qui explique cette faible différence de résultats. Par ailleurs, le graphical Lasso peut avoir tendance à gommer ces différences également. En effet, l'impact de l'estimation de ces points de troncature sera sûrement plus visible dans l'estimation de la matrice de covariance. Mais cet effet peut être par la suite "lissé" avec la pénalisation du graphical Lasso.

6.3.5 Impact des points de troncature

Pour compléter cette section, on propose d'illustrer l'impact des points de troncature sur l'estimation du réseau. Pour cela, on compare brièvement les résultats obtenus selon les configurations choisies pour les points de troncature : les observations du vecteur gaussien sont les mêmes et on change seulement les points de troncature pour pouvoir effectuer ces comparaisons. On applique notre procédure dans le cas où les points de troncature sont connus. Le paramètre de pénalisation est choisi avec la méthode "ebic".

Résultats et commentaires. La figure 6.6 représente les taux de détection des arêtes pour les quatre différentes configurations des points de troncature rappelées ci-dessous :

- seuils identiques : $a = -0.5$ et $b = 2$,
- seuils symétriques : $a = -1.5$ et $b = 1.5$,
- seuils croissants : $a =$ `seq(-1,0, length = p)`, $b =$ `seq(0.5,3, length = p)`,
- seuils décroissants : $a = -1$, $b =$ `seq(2,0.5, length = p)`.

Les configurations "seuils identiques" et "seuils symétriques" donnent des résultats assez similaires. La première configuration donne une inflation de zéros d'environ 33% contre 13% pour la deuxième dans les données Y. Assez étonnamment, les vraies arêtes semblent légèrement mieux détectées pour la configuration "seuils identiques". Cependant, les fausses arêtes sont légèrement plus détectées également.
Les données de la configuration "seuils décroissants" comportent de 20% (pour Y_1) à 50% (pour Y_{100}) d'inflation de zéros. Les taux de détection des potentielles arêtes sont représentés en déroulant la matrice à l'aide de la commande `upper.tri` de R, c'est-à-dire que la première arête dont le taux de détection est représenté est l'arête $X_2 \longleftrightarrow X_1$, puis $X_3 \longleftrightarrow X_1$, $X_3 \longleftrightarrow X_2$, $X_4 \longleftrightarrow X_1$, $\ldots$, $X_4 \longleftrightarrow X_3$ etc. Les 99 vraies arêtes théoriques, c'est-à-dire les arêtes $X_i \longleftrightarrow X_{i+1}$, sont représentées en rouge et dans l'ordre $X_1 \longleftrightarrow X_2$, $X_2 \longleftrightarrow X_3$, $\ldots$, $X_{99} \longleftrightarrow X_{100}$. Ainsi, on peut constater que les arêtes impliquant des variables dont la proportion de zéros approche les 50% ont des taux de détection moins bons : les vraies arêtes sont moins bien détectées tandis que les fausses arêtes le sont davantage.
On constate le même genre de phénomène avec la configuration "seuils croissants". Cependant, dans cette configuration, l'inflation de zéros décroît de 47% (pour Y_1) à 34% (pour Y_{40}) et recroît jusqu'à 50% (pour Y_{100}). On observe ce phénomène sur les arêtes impliquant les premières variables (les vraies arêtes sont moins bien détectées tandis que les fausses le sont davantage) mais pas sur les arêtes impliquant les variables Y_i pour i proche de 100, ce qui est assez étonnant. Ceci peut être dû au fait que la troncature sur les dernières variables donne davantage d'informations sur ces gaussiennes (qui sont "observées" entre 0 et 3) que sur les premières variables (observées à leurs vraies valeurs entre -1 et 0.5). En d'autres termes, l'idée est que la qualité de l'observation d'une variable gaussienne centrée réduite est meilleure entre 0 et 3 qu'entre -1 et 0.5, notamment pour l'estimation des covariances. Cette explication est corroborée par la figure

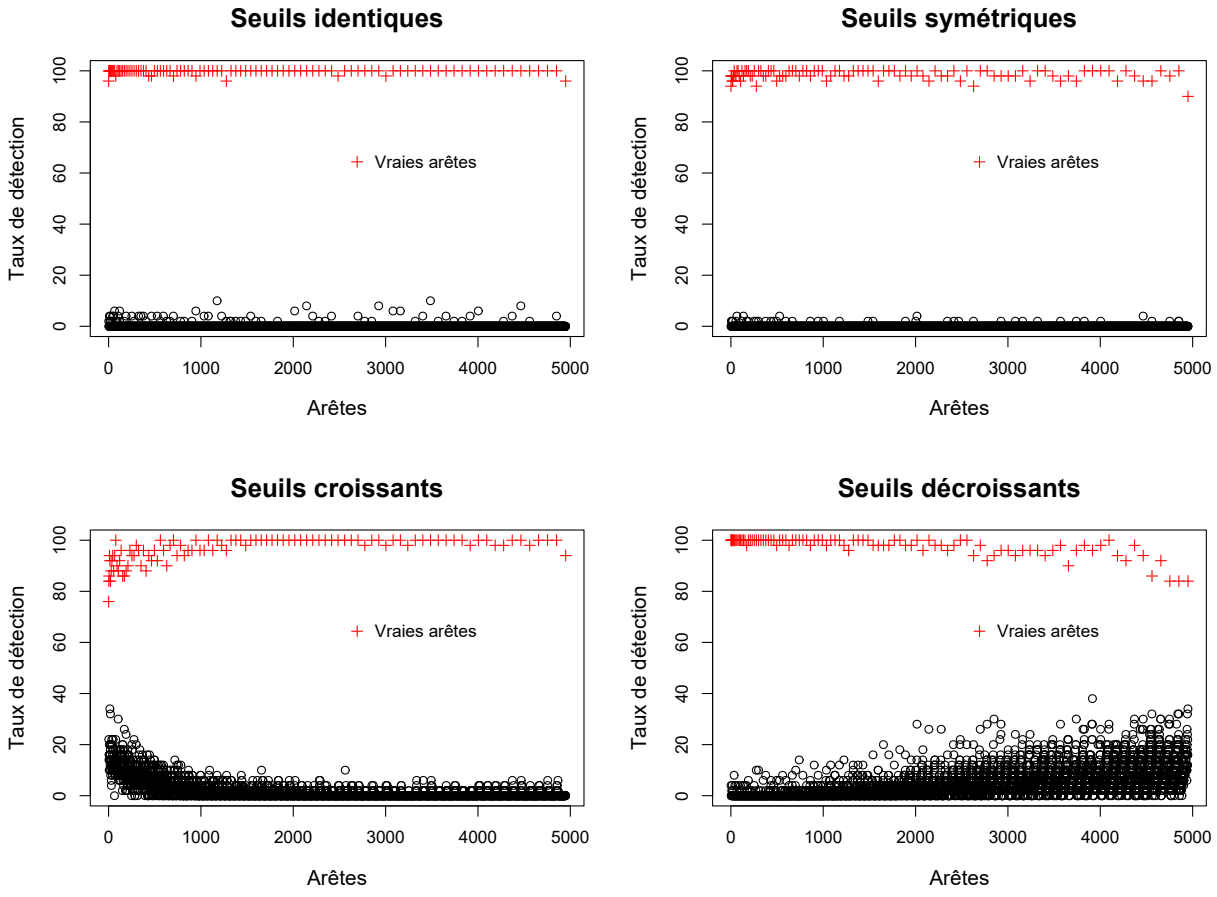


FIGURE 6.6 – Comparaison des taux de détection selon les configurations des points de troncature, quand ceux-ci sont connus. On compare les configurations “seuils identiques”, “seuils symétriques”, “seuils croissants” et “seuils décroissants”. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

6.7. Sur cette figure, on compare les résultats obtenus pour les seuils identiques, pour lesquels le taux d'inflation de zéros vaut 33%, avec ceux obtenus pour les seuils $a = -1$ et $b = 1$, pour lesquels le taux d'inflation de zéros est similaire et vaut 32%. Pourtant, les résultats obtenus pour les seuils identiques sont bien meilleurs, ce qui confirme l'hypothèse émise sur la fenêtre d'observation de la gaussienne.

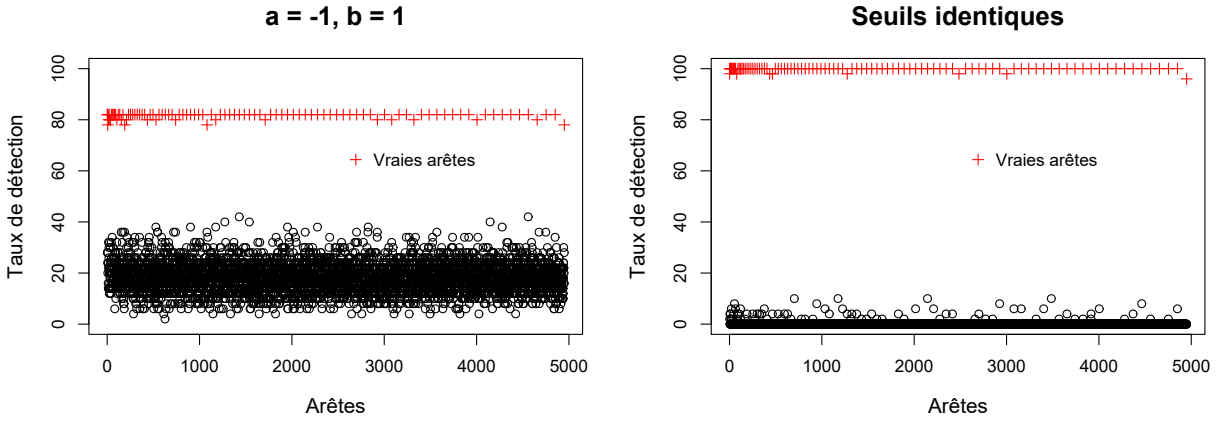


FIGURE 6.7 – Comparaison des taux de détection pour deux configurations des points de troncature pour lesquelles le taux d'inflation en zéro est similaire (environ 32%). Les points de troncature sont estimés. On compare la configuration "seuils identiques" ($a = -0.5$ et $b = 2$) avec la configuration $a = -1$ et $b = 1$. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

6.3.6 Autres structures de graphes

Jusqu'ici, on s'est limité à une structure de graphe "chaîne", qui a tendance à donner de bons résultats d'une manière générale. Ceci est en partie dû aux degrés constants des sommets du graphe (presque tous les sommets du graphe ont deux voisins). Pour compléter ces études de simulations, on présente ici des résultats sur d'autres structures de graphes :

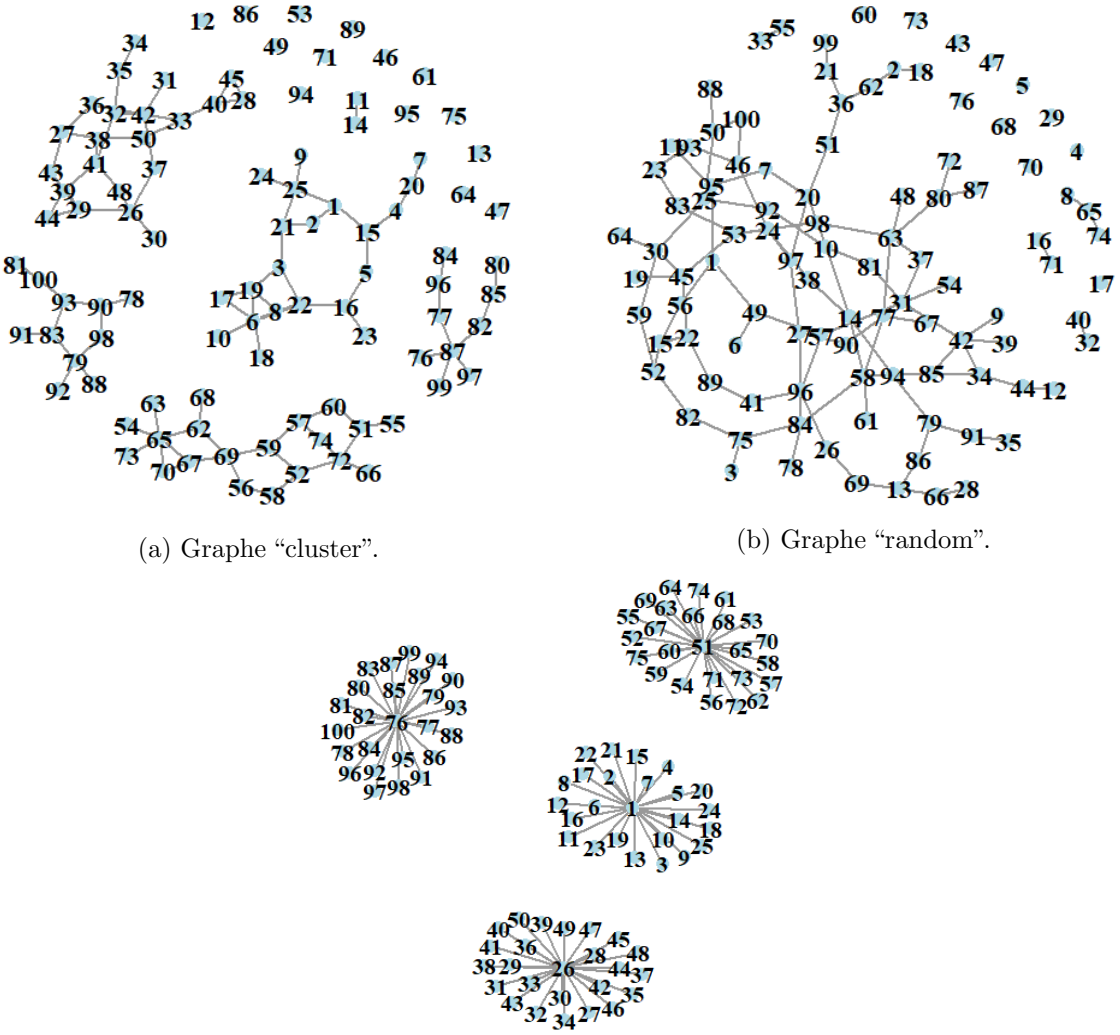
— La structure "cluster". Les variables sont découpées en 4 groupes de 25. Au sein d'un groupe, il existe une arête entre deux variables avec probabilité 0.0825. Il n'existe pas d'arêtes entre deux variables de deux groupes différents. Les données ont été simulées avec la fonction `huge.generator`, options `graph = 'cluster', g = 4, prob = 0.0825`. Le graphe résultant comporte 100 arêtes.

— La structure "random". Il existe une arête entre deux variables avec probabilité 1/50. Les données ont été simulées avec la fonction `huge.generator`, options `graph = 'random', prob = 1/50`. Le graphe résultant comporte 103 arêtes.

— La structure "hub". Les variables sont découpées en 4 groupes de 25. Au sein de chaque groupe, une des variables devient un "hub" et est connectée à toutes les variables de son groupe. Les données ont été simulées avec la fonction `huge.generator`, options `graph = 'hub', g = 4`. Le graphe résultant comporte 96 arêtes.

Ces graphes sont représentés à la figure 6.8. Concernant les seuils de troncature, on utilise la configuration "seuils identiques" ($a = -0.5$ et $b = 2$). On simule $n = 500$ observations de $p = 100$ variables et on compare les taux de détection obtenus avec notre procédure (points de troncature

estimés) à ceux obtenus avec le graphical Lasso directement sur les données tronquées, comme à la sous-section 6.3.3.



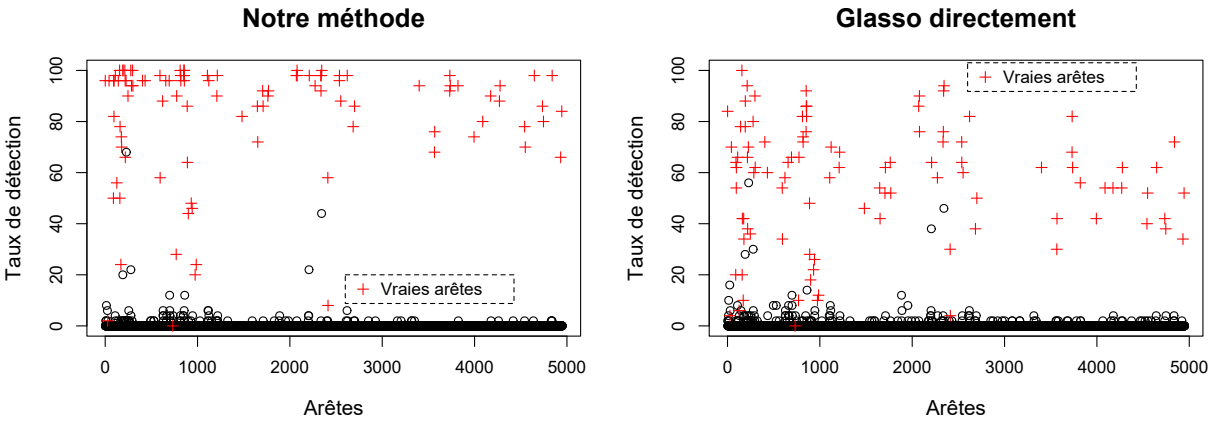
(a) Graphe “cluster”.

(b) Graphe “random”.

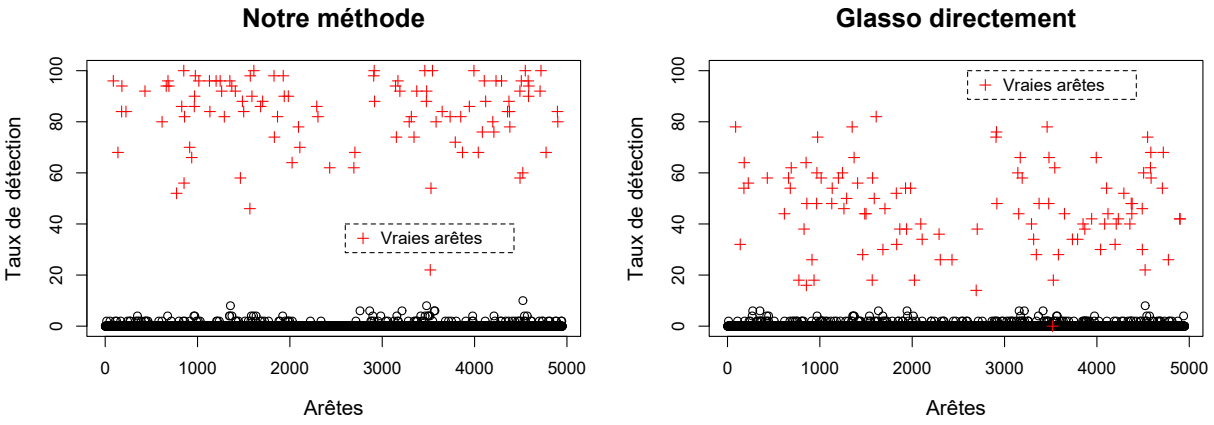
(c) Graphe “hub”.

FIGURE 6.8 – Représentation graphique des trois graphes utilisés dans cette sous-section : “cluster”, “random” et “hub”.

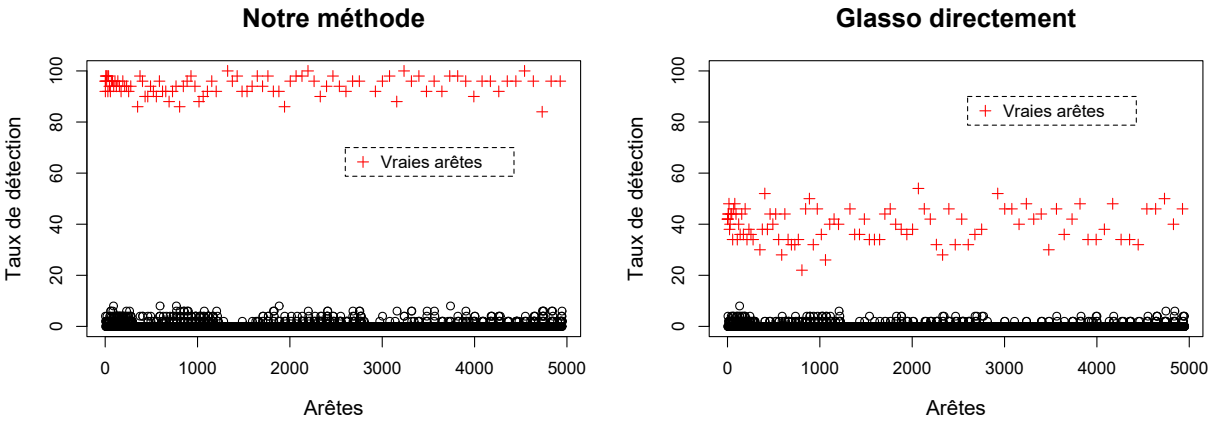
Résultats et commentaires. La figure 6.9 montre ces comparaisons pour les trois structures de graphes “cluster”, “random” et “hub”.
Notre procédure donne en effet des résultats globalement moins bons que sur la structure “chaîne”. Pour la structure “cluster”, 30 vraies arêtes sont détectées moins de 60% dont respectivement 7 et 4 moins de 28% et 20%. Il s’agit des arêtes $X_{29} \longleftrightarrow X_{39}$ (0%) $X_6 \longleftrightarrow X_8$ (2%), $X_{67} \longleftrightarrow X_{70}$ (8%), $X_{28} \longleftrightarrow X_{45}$ (20%), $X_{17} \longleftrightarrow X_{19}$ (24%), $X_{40} \longleftrightarrow X_{45}$ (24%) et $X_{28} \longleftrightarrow X_{40}$ (28%). En termes de faux positifs, 5 arêtes sont détectées à tort plus de 20% : les arêtes $X_{19} \longleftrightarrow X_{22}$ (68%), $X_{65} \longleftrightarrow X_{69}$ (44%), $X_2 \longleftrightarrow X_{25}$ et $X_{62} \longleftrightarrow X_{67}$ (22%) et $X_1 \longleftrightarrow X_{21}$ (20%). Au regard de la figure 6.8a, on constate que ces 5 couples de variables sont à chaque fois indirectement liés par deux autres variables (X_3 et X_8 ; X_{62} et X_{67} ; X_2 et X_{25} ;



(a) Structure “cluster”.



(b) Structure “random”.



(c) Structure “hub”.

FIGURE 6.9 – Comparaison des taux de détection obtenus par notre méthode (points de troncature estimés) et par le graphical Lasso appliqué directement sur les données tronquées. Les configurations des points de troncature sont “seuils identiques”. Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables. Structures “cluster”, “random” et “hub” ; en rouge sont représentées les vraies arêtes.

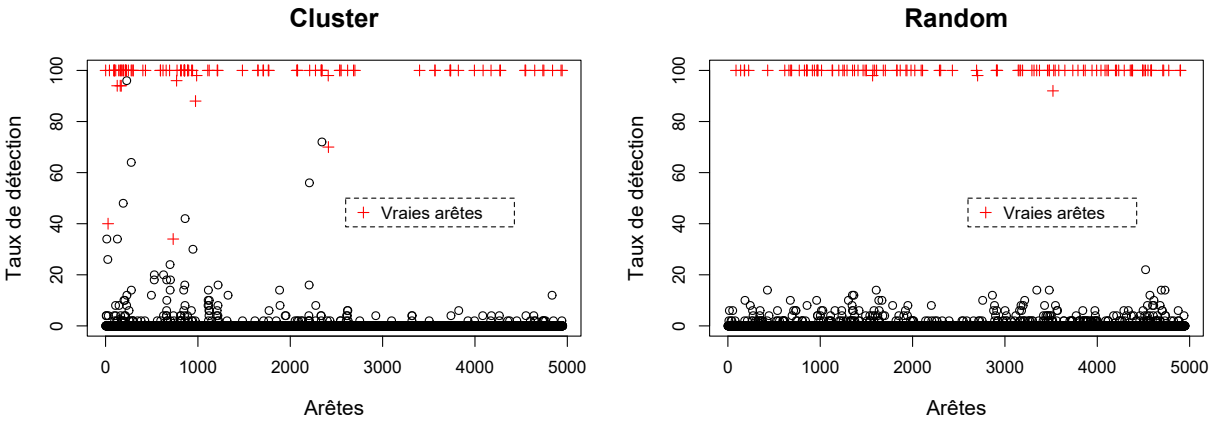


FIGURE 6.10 – Taux de détection obtenus avec le Glasso sur les données gaussiennes non tronquées pour les structures "cluster" et "random". Les taux de détection sont obtenus sur 50 répétitions indépendantes pour $n = 500$ observations de $p = 100$ variables.

X_{65} et X_{69} ; X_2 et X_{25}). Sur la figure 6.10, on peut en fait constater que la détection des arêtes n'est pas très bonne même en appliquant directement le Glasso sur les données gaussiennes non tronquées (en comparaison avec la structure "random" notamment). Ces problèmes de détection semblent donc liés directement à la structure de graphe et pas forcément à la troncature.
Les taux de détection de la structure "random" sont un peu meilleurs. Toutes les fausses arêtes sont détectées moins de 10% et les vraies arêtes sont toutes détectées plus de 46% sauf l'arête $X_{34} \longleftrightarrow X_{85}$ (22%). Cette arête est également légèrement moins détectée (92% contre au moins 98% pour les autres vraies arêtes) quand on applique le Glasso directement sur les données gaussiennes non tronquées (voir figure 6.10).
La stucture "hub" donne de bons résultats : les fausses arêtes sont détectées au plus 8% et les vraies arêtes au moins 84%.
En comparaison, Glasso donne toujours de moins bons résultats. Pour "cluster", les mêmes fausses arêtes sont beaucoup détectées et les vraies arêtes sont toutes moins bien détectées. Pour "random", l'arête $X_{34} \longleftrightarrow X_{85}$ n'est jamais détectée et les autres vraies arêtes sont détectées entre 14% et 82%. La structure "hub" montre la différence la plus marquante : les vraies arêtes sont détectées au plus 54% (mais au moins 22%).

6.4 Conclusions

Dans ce chapitre, on a exposé une procédure pour l'inférence de réseaux dans un modèle où les données sont gaussiennes et inflatées en zéro par double troncature à droite et à gauche. Plus spécifiquement, l'objectif est de retrouver la structure latente de graphe (donnée par la matrice de précision du modèle gaussien) à partir des données tronquées. Cette procédure comporte deux étapes. La première consiste à estimer la matrice de covariance termes à termes par maximisation de la log-vraisemblance des couples de variables. La seconde étape s'appuie sur le graphical Lasso et a pour but de retrouver la structure de graphe en estimant la matrice de précision avec une pénalisation Lasso. Par la suite, on a montré que cette procédure présente des garanties théoriques asymptotiques intéressantes en termes d'estimation du graphe. Le résultat de l'article de Ravikumar *et al.* (2011) [83] sur l'estimation du graphe concerne en fait

l'estimation de la matrice de précision dans un cadre plus général. On pourrait alors appliquer leur résultat dans un modèle plus général où le modèle sous-jacent n'est pas forcément gaussien. En revanche, dans un modèle différent, la matrice de précision caractérise rarement la structure du graphe.
Au vu des simulations, cette procédure semble également pertinente d'un point de vue pratique et donne de meilleurs résultats que le graphical Lasso utilisé directement sur les données tronquées (sans une meilleure estimation préliminaire de la matrice de covariance). Il faut bien avoir en tête que ces résultats dépendent de la structure de graphe et les simulations montrent bien ce phénomène, certaines structures de graphe donnant, d'une manière générale, des résultats plus concordants quant à l'inférence.

Par la suite, il pourrait être intéressant d'explorer d'autres pistes pour la procédure d'estimation. On peut par exemple penser à estimer la matrice de covariance en un seul temps (au lieu de l'estimer termes à termes) en utilisant les vraisemblances composites, pour lesquelles la littérature comporte un certain nombre de résultats théoriques. On peut également penser à une procédure qui s'appuie sur l'algorithme EM et pour lequel les résultats théoriques sur l'estimateur du maximum de vraisemblance pourraient s'appliquer.
En outre, on se limite dans ce chapitre à une troncature bilatérale (à droite et à gauche). Les preuves de nos résultats théoriques requièrent ces deux bornes et ne fonctionnent plus dans le cas d'une troncature unilatérale. Toutefois, notre procédure semble fonctionner dans le cas unilatéral et une extension des résultats théoriques à ce cas serait sûrement possible avec d'autres outils.
Par ailleurs, les autres idées de procédures évoquées ci-dessus pourraient éventuellement aussi s'appliquer dans le cas d'une troncature unilatérale.

Conclusion générale et perspectives

Dans cette thèse, deux principaux axes ont été développés. Le premier concerne des méthodes d'inférence de réseaux basées sur l'estimation de voisinages. Ce type de méthode nous a amenés à considérer des méthodes de régressions pour des données ordinales et à développer une méthode de sélection de variables, la méthode des knockoffs revisités. Cet axe comporte beaucoup de travaux algorithmiques et notre méthode des knockoffs revisités pour la sélection de variables a notamment été implémentée dans un package R, `kosel`. Cependant, même si les résultats de simulations ont donné des résultats très prometteurs, les fondements théoriques de ces résultats algorithmiques n'ont pas été étudiés dans ce manuscrit et ceci pourrait faire l'objet de futurs travaux. Parmi eux, il serait notamment intéressant de développer les garanties théoriques de notre méthode de sélection de variables.
D'un point de vue algorithmique, ce sujet ayant été initialement motivé par une collaboration avec des biologistes de l'INRA, il pourrait être utile de développer un package à destination des biologistes pour l'inférence de réseaux. Ce package s'appuierait sur le package `kosel` et permettrait d'automatiser la méthode d'inférence basée sur l'estimation des voisinages développée dans le chapitre 5.

Le second axe traite d'inférence de réseaux dans un modèle gaussien inflaté en zéro par double troncature, appelée aussi troncature bilatérale. Cette seconde partie présente davantage d'aspects théoriques que nous pourrions toutefois accroître et étendre à un modèle similaire où la troncature est unilatérale. L'extension de ces résultats peut être menée en considérant une autre procédure d'estimation basée sur des outils différents comme les vraisemblances composites ou l'algorithme EM. Par ailleurs, nous pourrions implémenter la procédure d'estimation proposée dans ce modèle dans un package et il serait intéressant d'étudier son application sur des données réelles.

Ces deux axes restent cependant très spécifiques et ne constituent qu'une contribution partielle à ce sujet. Même si les particularités des jeux de données inflatés en zéro et les contraintes liées à leur structure de dépendance conditionnelle rendent leur modélisation complexe, les modèles proposés dans ce manuscrit sont loin d'être les seuls adaptés. Il serait intéressant de réfléchir à de nouveaux modèles répondant à ces critères, qui permettraient également de modéliser de façon plus pertinente les effets de la réplication et des processus de séquençage utilisés par les biologistes. Par exemple, un modèle plus sophistiqué mélangeant le modèle gaussien et le modèle d'Ising pourrait être une piste, dans la mesure où la structure de dépendance conditionnelle est connue théoriquement pour ces deux modèles.

Annexe A

Preuves et compléments du chapitre 6

Démonstration. (Lemme 6.2.1) Il suffit de montrer l'existence d'une telle constante $\gamma_{jk} > 0$ pour $j < k$ fixés et de prendre $\gamma = \max_{j<k} \gamma_{jk} > 0$. Soient $j < k$, $\sigma \in [-1+\delta, 1-\delta]$ et $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$.

- $a = b = 1$:

$$\begin{aligned}\phi_{11,jk}(\sigma, y_j, y_k) &= f(y_j, y_k, \sigma) \\ &= \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\Big[- \frac{y_j^2 - 2\sigma y_j y_k + y_k^2}{2(1-\sigma^2)}\Big].\end{aligned}$$

Comme $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ et $\delta^2 \leq 1 - \sigma^2 \leq 1$, $\phi_{11,jk}$ est continue et strictement positive sur un compact de $\mathbb{R}^3$ et on a le résultat voulu.

- $a = 0, b = 1$:

$$\begin{aligned}\phi_{01,jk}(\sigma, y_j, y_k) &= \phi_{01,jk}(\sigma, y_k) \\ &= \int_{[a_j,b_j]^c} f(x, y_k, \sigma)\mathrm{d}x \\ &= \frac{\exp\Big(- \dfrac{y_k^2}{2(1-\sigma^2)}\Big)}{2\pi\sqrt{1-\sigma^2}} \int_{[a_j,b_j]^c} \exp\Big(- \frac{(x-\sigma y_k)^2}{2(1-\sigma^2)}\Big) \exp\Big(\frac{\sigma^2 y_k^2}{2(1-\sigma^2)}\Big)\mathrm{d}x \\ &= \frac{\exp\Big(- \dfrac{y_k^2}{2}\Big)}{2\pi\sqrt{1-\sigma^2}} \int_{\left[\frac{a_j-\sigma y_k}{\sqrt{1-\sigma^2}}, \frac{b_j-\sigma y_k}{\sqrt{1-\sigma^2}}\right]^c} \sqrt{1-\sigma^2} \exp\Big(- \frac{x^2}{2}\Big)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{y_k^2}{2}\Big)\Big[1 - \Phi\big(\frac{b_j - \sigma y_k}{\sqrt{1-\sigma^2}}\big) + \Phi\big(\frac{a_j - \sigma y_k}{\sqrt{1-\sigma^2}}\big)\Big], \qquad (A.1)\end{aligned}$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$. Comme $y_k \in [a_k, b_k]$, $-\infty < a_k < b_k < \infty$ et $\delta^2 \leq 1 - \sigma^2 \leq 1$, $\phi_{01,jk}$ est donc continue et strictement positive sur un compact de $\mathbb{R}^2$ et on a le résultat voulu.

- $a = 1, b = 0$: idem que pour $a = 0, b = 1$.

- $a = b = 0$:

$$\begin{aligned}\phi_{00,jk}(\sigma, y_j, y_k) &= \phi_{00,jk}(\sigma) \\ &= \iint_{[a_j,b_j]^c\times[a_k,b_k]^c} f(x,y,\sigma)\mathrm{d}x\mathrm{d}y \\ &= \int_{[a_k,b_k]^c} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2}\Big)\Big[1-\Phi\Big(\frac{b_j-\sigma y}{\sqrt{1-\sigma^2}}\Big)+\Phi\Big(\frac{a_j-\sigma y}{\sqrt{1-\sigma^2}}\Big)\Big]\mathrm{d}y.\end{aligned}$$

Comme $\delta^2 \leq 1-\sigma^2 \leq 1$, $\phi_{00,jk}$ est donc continue et strictement positive sur un compact de $\mathbb{R}$ et on a le résultat voulu. □

Démonstration. (Lemme 6.2.2) Comme pour la démonstration du lemme 6.2.1, il suffit de montrer ce résultat pour $j < k$ fixés. Soit $j < k$ et montrons que pour tout $a, b \in \{0,1\}$, la fonction $\phi_{ab,jk}$ est C^3 sur $[-1+\delta, 1-\delta]\times[a_j,b_j]\times[a_k,b_k]$ (en fait, elle y est même C^∞). Ainsi, pour tout $m \in \{1,2,3\}$, $\partial_\sigma^m \phi_{ab,jk}$ est continue sur un compact de $\mathbb{R}^3$, ce qui assure le résultat cherché.

- $a = b = 1$:

$$\phi_{11,jk}(\sigma, y_j, y_k) = \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\Big[-\frac{y_j^2 - 2\sigma y_j y_k + y_k^2}{2(1-\sigma^2)}\Big].$$

$\phi_{11,jk}$ est C^3 sur $]-1,1[\times\mathbb{R}^2$ donc sur $[-1+\delta,1-\delta]\times[a_j,b_j]\times[a_k,b_k]$.

- $a = 0, b = 1$:

$$\phi_{01,jk}(\sigma, y_j, y_k) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y_k^2}{2}\Big)\Big[1-\Phi\Big(\frac{b_j-\sigma y_k}{\sqrt{1-\sigma^2}}\Big)+\Phi\Big(\frac{a_j-\sigma y_k}{\sqrt{1-\sigma^2}}\Big)\Big],$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$ (voir (A.1)). Comme Φ est C^∞ donc C^3 sur $\mathbb{R}$, $\phi_{01,jk}$ est C^3 sur $]-1,1[\times\mathbb{R}$ donc sur $[-1+\delta,1-\delta]\times[a_k,b_k]$.

- $a = 1, b = 0$: idem que pour $a = 0, b = 1$.

- $a = b = 0$:

$$\begin{aligned}\phi_{00,jk}(\sigma, y_j, y_k) &= \int_{[a_k,b_k]^c} \phi_{01,jk}(\sigma, y)\mathrm{d}y \\ &= \int_{[a_k,b_k]^c} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2}\Big)\Big[1-\Phi\Big(\frac{b_j-\sigma y}{\sqrt{1-\sigma^2}}\Big)+\Phi\Big(\frac{a_j-\sigma y}{\sqrt{1-\sigma^2}}\Big)\Big]\mathrm{d}y.\end{aligned}$$

Soit $m \in \{1,2,3\}$. On va appliquer les théorèmes de continuité et dérivation des intégrales à paramètres :

— $\sigma \in [-1+\delta, 1-\delta] \mapsto \phi_{01,jk}(\sigma, y)$ est C^3.

— En posant $h_a = \dfrac{a - \sigma y_k}{\sqrt{1-\sigma^2}}$, on a :

$$\begin{aligned}
\partial_\sigma \phi_{01,jk}(\sigma, y) &= \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2}\Big)\Big[- h'_{b_j}\Phi'\Big(h_{b_j}\Big) + h'_{a_j}\Phi'\Big(h_{a_j}\Big)\Big)\Big], \\
\partial^2_\sigma \phi_{01,jk}(\sigma, y) &= \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y^2}{2}\Big)\Big[- h^{(2)}_{b_j}\Phi'\Big(h_{b_j}\Big) - h'^2_{b_j}\Phi^{(2)}\Big(h_{b_j}\Big) \\
&\qquad + h^{(2)}_{a_j}\Phi'\Big(h_{a_j}\Big) + h'^2_{a_j}\Phi^{(2)}\Big(h_{a_j}\Big)\Big], \\
\partial^3_\sigma \phi_{01,jk}(\sigma, y) &= \frac{\exp\Big(-\dfrac{y^2}{2}\Big)}{\sqrt{2\pi}}\Big[- h^{(3)}_{b_j}\Phi'\Big(h_{b_j}\Big) - 3h'_{b_j}h^{(2)}_{b_j}\Phi^{(2)}\Big(h_{b_j}\Big) - h'^3_{b_j}\Phi^{(3)}\Big(h_{b_j}\Big) \\
&\qquad + h^{(3)}_{a_j}\Phi'\Big(h_{a_j}\Big) + 3h'_{a_j}h^{(2)}_{a_j}\Phi^{(2)}\Big(h_{a_j}\Big) + h'^3_{a_j}\Phi^{(3)}\Big(h_{a_j}\Big)\Big],
\end{aligned}$$

où $h^{(r)}$ désigne la dérivée r-ième de h par rapport à σ. $h^{(r)}$ est une somme de termes de la forme $\dfrac{Cy^{r_1}\sigma^{r_2}}{\sqrt{(1-\sigma^2)^{r_3}}}$ où $r_1, r_2, r_3 \in \mathbb{N}$ et $C \in \mathbb{R}$.

— Pour $\sigma \in [-1+\delta, 1-\delta]$:

$$\Big|\partial^m_\sigma \phi_{01,jk}(\sigma, y)\Big| \leq C(a_j, b_j, a_k, b_k, \delta, m) \exp\Big(-\frac{y^2}{2}\Big),$$

où $C(a_j, b_j, a_k, b_k, \delta, m)$ est une constante positive dépendant de $a_j, b_j, a_k, b_k, \delta$ et de m et $y \mapsto C(a_j, b_j, a_k, b_k, \delta, m) \exp\Big(-\dfrac{y^2}{2}\Big)$ est intégrable sur $[a_k, b_k]^c$.

On en déduit que $\phi_{00,jk}$ est C^3 sur $[-1+\delta, 1-\delta]$. □

Démonstration. (Lemme 6.2.3)

1. Soit $l \in \mathbb{N}^*$. Tout d'abord, on a, pour tout $\sigma \in [-1+\delta, 1-\delta]$:

$$\left.\begin{aligned}
\int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) &= \phi_{00,jk}(\sigma) + \int_{a_k}^{b_k} \phi_{01,jk}(\sigma, y)\mathrm{d}y + \int_{a_j}^{b_j} \phi_{10,jk}(\sigma, x)\mathrm{d}x \\
&\quad + \iint_{[a_j,b_j]\times[a_k,b_k]} \phi_{11,jk}(\sigma, x, y)\mathrm{d}x\mathrm{d}y \\
&= \iint_{\mathbb{R}^2} f(x, y, \sigma)\mathrm{d}x\mathrm{d}y = 1.
\end{aligned}\right\} \tag{A.2}$$

D'une part, on a bien :

$$\partial^l_\sigma\Big(\int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y)\Big) = \partial^l_\sigma(1) = 0.$$

D'autre part, montrons que :

$$\partial_\sigma^l\Big(\int_{\mathbb{R}^2}\mathcal{L}_{jk}(\sigma,y)\mathrm{d}\mu(y)\Big)=\partial_\sigma^l\phi_{00,jk}(\sigma)+\int_{a_k}^{b_k}\partial_\sigma^l\phi_{01,jk}(\sigma,y)\mathrm{d}y+\int_{a_j}^{b_j}\partial_\sigma^l\phi_{10,jk}(\sigma,x)\mathrm{d}x \tag{A.3}$$
$$+\iint_{[a_j,b_j]\times[a_k,b_k]}\partial_\sigma^l\phi_{11,jk}(\sigma,x,y)\mathrm{d}x\mathrm{d}y$$
$$=\int_{\mathbb{R}^2}\partial_\sigma^l\mathcal{L}_{jk}(\sigma,y)\mathrm{d}\mu(y).$$

Pour montrer (A.3), il suffit de montrer que c'est vrai pour chacun des quatre termes :

• Pour $a=0,b=0$: c'est clair.
Pour les termes suivants, on va utiliser le théorème de dérivation sous le signe somme.

• Pour $a=0,b=1$ (et $a=1,b=0$) :

$$\phi_{01,jk}(\sigma,y_j,y_k)=\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{y_k^2}{2}\Big)\Big[1-\Phi\Big(\frac{b_j-\sigma y_k}{\sqrt{1-\sigma^2}}\Big)+\Phi\Big(\frac{a_j-\sigma y_k}{\sqrt{1-\sigma^2}}\Big)\Big],$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$. La fonction $\phi_{01,jk}$ est C^∞ sur $[1-\delta,1+\delta]\times[a_k,b_k]$. Ainsi, pour tout $\sigma\in[-1+\delta,1-\delta]$ et $y\in[a_k,b_k]$, $\partial_\sigma^l\phi_{01,jk}(\sigma,y)$ est borné par une constante, qui est donc intégrable sur le compact $[a_k,b_k]$. Ainsi, on a bien :

$$\partial_\sigma^l\Big(\int_{a_k}^{b_k}\phi_{01,jk}(\sigma,y)\mathrm{d}y\Big)=\int_{a_k}^{b_k}\partial_\sigma^l\phi_{01,jk}(\sigma,y)\mathrm{d}y.$$

• Pour $a=1,b=1$:

$$\phi_{11,jk}(\sigma,x,y)=\frac{1}{2\pi\sqrt{1-\sigma^2}}\exp\Big[-\frac{x^2-2\sigma xy+y^2}{2(1-\sigma^2)}\Big].$$

De même, $\phi_{11,jk}$ est C^∞ sur $[1-\delta,1+\delta]\times[a_j,b_j]\times[a_k,b_k]$. Ainsi, pour tout $\sigma\in[-1+\delta,1-\delta]$, $x\in[a_j,b_j]$ et $y\in[a_k,b_k]$, $\partial_\sigma^l\phi_{11,jk}(\sigma,x,y)$ est borné par une constante, qui est donc intégrable sur le compact $[a_j,b_j]\times[a_k,b_k]$. Ainsi, on a bien :

$$\partial_\sigma^l\Big(\iint_{[a_j,b_j]\times[a_k,b_k]}\phi_{11,jk}(\sigma,x,y)\mathrm{d}x\mathrm{d}y\Big)=\iint_{[a_j,b_j]\times[a_k,b_k]}\partial_\sigma^l\phi_{11,jk}(\sigma,x,y)\mathrm{d}x\mathrm{d}y.$$

2. Soit $l\in\mathbb{N}^*$. Explicitons d'abord quelques notations :

$$\mathbb{E}_{\Sigma_{jk}^*}\Big(\log\mathcal{L}_{jk}(\sigma,Y)\Big)=\mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma,Y)\Big)=R(\sigma)$$
$$=\int_{\mathbb{R}^2}\log\mathcal{L}_{jk}(\sigma,y)\mathcal{L}_{jk}(\Sigma_{jk}^*,y)\mathrm{d}\mu(y)$$
$$=\phi_{00,jk}(\Sigma_{jk}^*)\log\phi_{00,jk}(\sigma)+\int_{a_k}^{b_k}\phi_{01,jk}(\Sigma_{jk}^*,y)\log\phi_{01,jk}(\sigma,y)\mathrm{d}y$$
$$+\int_{a_j}^{b_j}\phi_{10,jk}(\Sigma_{jk}^*,x)\log\phi_{10,jk}(\sigma,x)\mathrm{d}x$$
$$+\iint_{[a_j,b_j]\times[a_k,b_k]}\phi_{11,jk}(\Sigma_{jk}^*,x,y)\log\phi_{11,jk}(\sigma,x,y)\mathrm{d}x\mathrm{d}y.$$

D'autre part,

$$\begin{aligned}\mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^l_\sigma \log \mathcal{L}_{jk}(\sigma, Y)\Big) &= \int_{\mathbb{R}^2} \partial^l_\sigma\Big(\log \mathcal{L}_{jk}(\sigma, y)\Big)\mathcal{L}_{jk}(\Sigma^*_{jk}, y)\mathrm{d}\mu(y)\\ &= \phi_{00,jk}(\Sigma^*_{jk})\partial^l_\sigma \log \phi_{00,jk}(\sigma) + \int_{a_k}^{b_k} \phi_{01,jk}(\Sigma^*_{jk}, y)\partial^l_\sigma \log \phi_{01,jk}(\sigma, y)\mathrm{d}y\\ &\quad + \int_{a_j}^{b_j} \phi_{10,jk}(\Sigma^*_{jk}, x)\partial^l_\sigma \log \phi_{10,jk}(\sigma, x)\mathrm{d}x\\ &\quad + \iint_{[a_j,b_j]\times[a_k,b_k]} \phi_{11,jk}(\Sigma^*_{jk}, x, y)\partial^l_\sigma \log \phi_{11,jk}(\sigma, x, y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

Ainsi, pour montrer l'égalité $\partial^l_\sigma \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) = \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^l_\sigma \log \mathcal{L}_{jk}(\sigma, Y)\Big)$, on va montrer l'égalité pour chacun des quatre termes.

• Pour $a = 0, b = 0$: c'est clair.
Pour les trois termes suivants, on va utiliser le théorème de dérivation sous le signe somme.

• Pour $a = 0, b = 1$ (et $a = 1, b = 0$) :

$$\log \phi_{01,jk}(\sigma, y) = -\frac{y^2}{2} - \log\sqrt{2\pi} + \log\left[1 - \Phi\Big(\frac{b_j - \sigma y}{\sqrt{1-\sigma^2}}\Big) + \Phi\Big(\frac{a_j - \sigma y}{\sqrt{1-\sigma^2}}\Big)\right].$$

La fonction $\log \phi_{01,jk}$ est C^∞ sur $[-1+\delta, 1-\delta]\times[a_k, b_k]$. Ainsi, pour tout $\sigma \in [-1+\delta, 1-\delta]$ et $y \in [a_k, b_k]$, $\left|\partial^l_\sigma \log \phi_{01,jk}(\sigma, y)\right|$ est majoré par une constante qui est intégrable sur le compact $[a_k, b_k]$ (par rapport à la densité $y \mapsto \phi_{01,jk}(\Sigma^*_{jk}, y)$). On a alors bien :

$$\partial^l_\sigma \int_{a_k}^{b_k} \phi_{01,jk}(\Sigma^*_{jk}, y)\log \phi_{01,jk}(\sigma, y)\mathrm{d}y = \int_{a_k}^{b_k} \phi_{01,jk}(\Sigma^*_{jk}, y)\partial^l_\sigma \log \phi_{01,jk}(\sigma, y)\mathrm{d}y.$$

• Pour $a = 1, b = 1$:

$$\log \phi_{11,jk}(\sigma, x, y) = -\log(2\pi) - \frac{1}{2}\log(1-\sigma^2) - \frac{x^2 - 2\sigma xy + y^2}{2(1-\sigma^2)}.$$

La fonction $\log \phi_{11,jk}$ est C^∞ sur $[-1+\delta, 1-\delta]\times[a_j, b_j]\times[a_k, b_k]$. Ainsi, pour tout $\sigma \in [-1+\delta, 1-\delta]$, $x \in [a_j, b_j]$ et $y \in [a_k, b_k]$, $\left|\partial^l_\sigma \log \phi_{11,jk}(\sigma, x, y)\right|$ est majoré par une constante qui est intégrable sur le compact $[a_j, b_j]\times[a_k, b_k]$ (par rapport à la densité $(x, y) \mapsto \phi_{11,jk}(\Sigma^*_{jk}, x, y)$). On a alors bien :

$$\begin{aligned}&\partial^l_\sigma \iint_{[a_j,b_j]\times[a_k,b_k]} \phi_{11,jk}(\Sigma^*_{jk}, x, y)\log \phi_{11,jk}(\sigma, x, y)\mathrm{d}x\mathrm{d}y\\ =&\iint_{[a_j,b_j]\times[a_k,b_k]} \phi_{11,jk}(\Sigma^*_{jk}, x, y)\partial^l_\sigma \log \phi_{11,jk}(\sigma, x, y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

□

Remarque : À l'aide du lemme 6.2.3, on peut également montrer facilement que Σ^*_{jk} est un

point critique de R. En effet,

$$\begin{aligned} R'(\Sigma^*_{jk}) &= \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial_\sigma \log \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)\Big) \text{ d'après le point 2} \\ &= \mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\partial_\sigma \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}\right) \\ &= \int_{\mathbb{R}^2} \partial_\sigma \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) \\ &= \partial_\sigma \int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) = 0 \text{ d'après le point 1.} \end{aligned}$$

Lemme A.0.1. *Soit X une variable aléatoire centrée et presque sûrement bornée par une constante $b > 0$, alors X est b^2-sous-gaussienne :*

$$\mathbb{E}\big(\exp(tX)\big) \leq \exp\Big(\frac{b^2t^2}{2}\Big) \text{ pour tout } t \in \mathbb{R}.$$

Démonstration. (Lemme A.0.1) On le montre d'abord pour $b = 1$. On considère, pour tout $t \in \mathbb{R}$, $f(t) := e^t\big(cosh(t) - \mathbb{E}(e^{tX})\big)$ et on veut montrer que f est positive.
On a en fait : $f(t) = \frac{1}{2}e^{2t} + \frac{1}{2} - \mathbb{E}(e^{t(X+1)})$. D'après le théorème de dérivation sous le signe somme, $f'(t) = e^{2t} - \mathbb{E}((X+1)e^{t(X+1)}) = \mathbb{E}\Big((X+1)(e^{2t} - e^{t(X+1)})\Big)$. Comme $0 \leq X + 1 \leq 2$ p.s., $(X+1)(e^{2t} - e^{t(X+1)}) \geq 0$ p.s. pour tout $t \geq 0$. Donc $f' \geq 0$ sur $\mathbb{R}^+$ et f est croissante sur $\mathbb{R}^+$. On a donc que pour tout $t \geq 0$, $f(t) \geq f(0) = 0$. Comme $-X$ est également centrée et presque sûrement bornée par 1, le résultat est aussi vrai pour $t \leq 0$. Ainsi :

$$\text{pour tout } t \in \mathbb{R},\ \mathbb{E}\big(\exp(tX)\big) \leq cosh(t) \leq \exp\Big(\frac{t^2}{2}\Big).$$

Si X est bornée par $b > 0$ quelconque, alors pour tout $t \in \mathbb{R}$:

$$\begin{aligned} \mathbb{E}\big(\exp(tX)\big) &= \mathbb{E}\big(\exp(bt\frac{X}{b})\big) \\ &\leq \exp\Big(\frac{b^2t^2}{2}\Big) \text{ d'après ce qui précède.} \end{aligned}$$

□

Bibliographie

[1] Alan Agresti. *Analysis of ordinal categorical data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1984.

[2] Alan Agresti. *Categorical data analysis*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1990. A Wiley-Interscience Publication.

[3] Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. Hiton : a novel markov blanket algorithm for optimal variable selection. In *AMIA Annual Symposium Proceedings*, volume 2003, page 21. American Medical Informatics Association, 2003.

[4] JA Anderson and PR Philips. Regression, discrimination and measurement models for ordered categorical variables. *Applied statistics*, pages 22–31, 1981.

[5] Ivan E. Auger and Charles E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1) :39–54, 1989.

[6] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9 :485–516, 2008.

[7] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5) :2055–2085, 2015.

[8] Rina Foygel Barber and Emmanuel J Candes. A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv :1602.03574*, 2016.

[9] Manjunath BG and Stefan Wilhelm. Moments calculation for the double truncated multivariate normal density. 2009.

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[11] Atul J Butte and Isaac S Kohane. Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[12] Atul J Butte and Isaac S Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*, page 711. American Medical Informatics Association, 1999.

[13] Frances Louise Campbell. *A study of truncated bivariate normal distributions*. ProQuest LLC, Ann Arbor, MI, 1945. Thesis (Ph.D.)–University of Michigan.

[14] Emmanuel Candès and Terence Tao. Rejoinder : "The Dantzig selector : statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351 ; mr2382644]. *Ann. Statist.*, 35(6) :2392–2404, 2007.

[15] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12) :4203–4215, 2005.

[16] Robert Castelo and Alberto Roverato. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, 7(Dec) :2621–2650, 2006.

[17] Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for sparse network reconstruction from count data. *arXiv preprint arXiv :1806.03120*, 2018.

[18] A. C. Cohen, Jr. On estimating the mean and standard deviation of truncated normal distributions. *J. Amer. Statist. Assoc.*, 44 :518–525, 1949.

[19] A. C. Cohen, Jr. Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *Ann. Math. Statistics*, 21 :557–569, 1950.

[20] A. Clifford Cohen. *Truncated and censored samples*, volume 119 of *Statistics : Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1991. Theory and applications.

[21] A. Clifford Cohen, Jr. Restriction and selection in samples from bivariate normal distributions. *J. Amer. Statist. Assoc.*, 50 :884–893, 1955.

[22] A. Clifford Cohen, Jr. On the solution of estimating equations for truncated and censored samples from normal populations. *Biometrika*, 44 :225–236, 1957.

[23] A. Clifford Cohen, Jr. Restriction and selection in multinormal distributions. *Ann. Math. Statist.*, 28 :731–741, 1957.

[24] D. R. Cox and Nanny Wermuth. *Multivariate dependencies*, volume 67 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. Models, analysis and interpretation.

[25] Harald Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

[26] Joachim Dahl, Lieven Vandenberghe, and Vwani Roychowdhury. Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software*, 23(4) :501–520, 2008.

[27] Bin Dai, Shilin Ding, and Grace Wahba. Multivariate Bernoulli distribution. *Bernoulli*, 19(4) :1465–1483, 2013.

[28] Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[29] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5) :2197–2202, 2003.

[30] Danny D. Dyer. On moments estimation of the parameters of a truncated bivariate normal distribution. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 22 :287–291, 1973.

[31] David Edwards. *Introduction to graphical modelling*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2000.

[32] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. With discussion, and a rejoinder by the authors.

[33] Karoline Faust and Jeroen Raes. Microbial interactions : from networks to models. *Nature Reviews Microbiology*, 10(8) :538, 2012.

[34] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

[35] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3 :95–110, 1956.

[36] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, 2008.

[37] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1, 2010.

[38] Carlo Gaetan and Xavier Guyon. *Modélisation et statistique spatiales*, volume 63 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2008.

[39] Christophe Giraud. Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2 :542–563, 2008.

[40] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.*, 11(3) :Art. 3, 52, 2012.

[41] A. K. Gupta and D. S. Tracy. Recurrence relations for the moments of truncated multinormal distribution. *Comm. Statist.–Theory Methods*, A5(9) :855–865, 1976.

[42] Peter Hall, Eun Ryung Lee, and Byeong U Park. Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica*, pages 449–471, 2009.

[43] David J. Hand. From evidence to understanding : a commentary on Fisher (1922) 'On the mathematical foundations of theoretical statistics'. *Philos. Trans. Roy. Soc. A*, 373(2039) :20140252, 8, 2015.

[44] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[45] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*, volume 143 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015. The lasso and generalizations.

[46] A Hoerl and R Kennard. Ridge regression, in 'encyclopedia of statistical sciences', vol. 8, 1988.

[47] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985.

[48] Jianhua Z. Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1) :85–98, 2006.

[49] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

[50] Boris Jakuschkin, Virgil Fievet, Loïc Schwaller, Thomas Fort, Cécile Robin, and Corinne Vacher. Deciphering the pathobiome : intra-and interkingdom interactions involving the pathogen erysiphe alphitoides. *Microbial ecology*, 72(4) :870–880, 2016.

[51] Ali Jalali, Christopher C Johnson, and Pradeep K Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems*, pages 1935–1943, 2011.

[52] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar) :613–636, 2007.

[53] Clémence Karmann and Aurélie Gueudin. Package `kosel`. `https://www.rdocumentation.org/packages/kosel`, 07/2019.

[54] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *J. Mach. Learn. Res.*, 8 :1519–1555, 2007.

[55] Ken Lang. Newsweeder : Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

[56] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

[57] Lung-fei Lee. The determination of moments of the doubly truncated multivariate normal tobit model. *Economics Letters*, 11(3) :245–250, 1983.

[58] Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of markov networks using l_1-regularization. In *Advances in neural Information processing systems*, pages 817–824, 2007.

[59] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient ℓ_1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

[60] Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Statist.*, 24(3) :627–654, 2015.

[61] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc., 2010.

[62] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization : An enabling technique. *Data mining and knowledge discovery*, 6(4) :393–423, 2002.

[63] Ivy Liu and Alan Agresti. The analysis of ordered categorical data : an overview and a survey of recent developments. *Test*, 14(1) :1–73, 2005. With discussion and a rejoinder by the authors.

[64] Po-Ling Loh and Martin J. Wainwright. Structure estimation for discrete graphical models : generalized covariance matrices and their inverses. *Ann. Statist.*, 41(6) :3022–3049, 2013.

[65] Dhafer Malouche and Sylvie Sevestre. Estimating high dimensional faithful Gaussian graphical models : upc-algorithm. Technical report, Citeseer, 2007.

[66] Peter McCullagh. Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B*, 42(2) :109–142, 1980.

[67] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[68] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv :1607.06534*, 2017.

[69] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1) :53–71, 2008.

[70] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1) :374–393, 2007.

[71] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3) :1436–1462, 2006.

[72] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4) :417–473, 2010.

[73] Bengt Muthén. Moments of the censored and truncated bivariate normal distribution. *British J. Math. Statist. Psych.*, 43(1) :131–143, 1990.

[74] G Baikunth Nath. Estimation in truncated bivariate normal distributions. *Applied Statistics*, pages 313–319, 1971.

[75] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4) :538–557, 2012.

[76] N.T.Viet and M.Paul. La simulation parfaite. `https://www.math.ens.fr/enseignement/telecharger_fichier.php?fichier=540`, 2007.

[77] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(4) :659–677, 2007.

[78] Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, Gilles Celeux, and Jean-Jacques Daudin. *A statistical approach for CGH microarray data analysis.* PhD thesis, INRIA, 2004.

[79] Franck Picard, Stéphane Robin, E Lebarbier, and J-J Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3) :758–766, 2007.

[80] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252, 1996.

[81] Des Raj. On estimating the parameters of bivariate normal populations from doubly and singly linearly truncated samples. *Sankhyā*, 12 :277–290, 1953.

[82] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3) :1287–1319, 2010.

[83] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5 :935–980, 2011.

[84] Vincent Rivoirard and Gilles Stoltz. *Statistique mathématique en action.* Vuibert, 2012.

[85] Laurent Rouvière. Régression logistique avec R. `https://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf`.

[86] Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6) :754–764, 2005.

[87] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4 :Art. 32, 28, 2005.

[88] S. M. Shah and M. C. Jaiswal. Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Ann. Inst. Statist. Math.*, 18 :107–111, 1966.

[89] Gary Simon. Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69(348) :971–976, 1974.

[90] Naunihal Singh. Estimation of parameters of a multivariate normal population from truncated and censored samples. *J. Roy. Statist. Soc. Ser. B*, 22 :307–311, 1960.

[91] Richard Splivallo, Maryam Vahdatzadeh, Jose Gaspar Maciá-Vicente, Virginie Molinier, Martina Peter, Simon Egli, Stephane Uroz, Francesco Paolocci, and Aurelie Deveau. Orchard conditions and fruiting body characteristics drive the microbiome of the black truffle tuber aestivum. *Frontiers in Microbiology*, 10 :1437, 2019.

[92] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.

[93] Arun Sai Suggala, Eunho Yang, and Pradeep Ravikumar. Ordinal graphical models : A tale of two approaches. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3260–3269, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[94] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996.

[95] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005.

[96] Joel A Tropp. Just relax : Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3) :1030–1051, 2006.

[97] Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection : relevancy, filters and wrappers. In *AISTATS*, 2003.

[98] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[99] Fanny Villers, Brigitte Schaeffer, Caroline Bertin, and Sylvie Huet. Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. *Stat. Appl. Genet. Mol. Biol.*, 7(1) :Art. 14, 36, 2008.

[100] Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2007.

[101] Strother H. Walker and David B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54 :167–179, 1967.

[102] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(3) :671–683, 2009.

[103] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A) :2178, 2009.

[104] Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3) :439–447, 1974.

[105] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7) :1669, 2016.

[106] Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1990.

[107] Wikistat. Introduction au modèle linéaire général — wikistat. `http://wikistat.fr/pdf/st-m-modlin-mlg.pdf`.

[108] Wikistat. An introduction to network inference and mining — wikistat. `http://www.nathalievilla.org/doc/pdf//wikistat-network_compiled.pdf`.

[109] Wikistat. Régression linéaire multiple ou modèle gaussien — wikistat. `http://wikistat.fr/pdf/st-m-modlin-regmult.pdf`.

[110] Wikistat. Régression logistique ou modèle binomial — wikistat. `http://wikistat.fr/pdf/st-m-app-rlogit.pdf`.

[111] Anja Wille and Peter Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, 5(1), 2006.

[112] O Dale Williams and James E Grizzle. Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association*, 67(337) :55–63, 1972.

[113] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6) :714–721, 2009.

[114] Michael J. Wurm, Paul J. Rathouz, and Bret M. Hanlon. Regularized Ordinal Regression and the ordinalNet R Package. *arXiv e-prints*, page arXiv :1706.05003, Jun 2017.

[115] Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

[116] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006.

[117] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1) :19–35, 2007.

[118] Peng Zhao, Guilherme Rocha, Bin Yu, et al. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A) :3468–3497, 2009.

[119] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 :2541–2563, 2006.

[120] Ji Zhu and Trevor Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3) :427–443, 2004.

[121] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476) :1418–1429, 2006.

[122] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005.

[123] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5) :2173–2192, 2007.

Résumé

L'inférence de réseaux ou inférence de graphes a de plus en plus d'applications notamment en santé humaine et en environnement pour l'étude de données micro-biologiques et génomiques. Les réseaux constituent en effet un outil approprié pour représenter, voire étudier des relations entre des entités. De nombreuses techniques mathématiques d'estimation ont été développées notamment dans le cadre des modèles graphiques gaussiens mais aussi dans le cas de données binaires ou mixtes.

Le traitement des données d'abondance (de micro-organismes comme les bactéries par exemple) est particulier pour deux raisons : d'une part elles ne reflètent pas directement la réalité car un processus de séquençage a lieu pour dupliquer les espèces et ce processus apporte de la variabilité, d'autre part une espèce peut être absente dans certains échantillons. On est alors dans le cadre de données inflatées en zéro. Beaucoup de méthodes d'inférence de réseaux existent pour les données gaussiennes, les données binaires et les données mixtes mais les modèles inflatés en zéro sont très peu étudiés alors qu'ils reflètent la structure de nombreux jeux de données de façon pertinente. L'objectif de cette thèse concerne l'inférence de réseaux pour les modèles inflatés en zéro.

Dans cette thèse, on se limitera à des réseaux de dépendances conditionnelles. Le travail présenté dans cette thèse se décompose principalement en deux parties. La première concerne des méthodes d'inférence de réseaux basées sur l'estimation de voisinages par une procédure couplant des méthodes de régressions ordinales et de sélection de variables. La seconde se focalise sur l'inférence de réseaux dans un modèle où les variables sont des gaussiennes inflatées en zéro par double troncature (à droite et à gauche).

Mots-clés: Inférence de graphes, réseaux, modèles inflatés en zéro, régression, pénalisation Lasso, sélection de variables, données gaussiennes doublement tronquées, dépendance conditionnelle.

Abstract

Network inference has more and more applications, particularly in human health and environment, for the study of micro-biological and genomic data. Networks are indeed an appropriate tool to represent, or even study, relationships between entities. Many mathematical estimation techniques have been developed, particularly in the context of Gaussian graphical models, but also in the case of binary or mixed data.

The processing of abundance data (of microorganisms such as bacteria for example) is particular for two reasons : on the one hand they do not directly reflect reality because a sequencing process takes place to duplicate species and this process brings variability, on the other hand a species may be absent in some samples. We are then in the context of zero-inflated data. Many graph inference methods exist for Gaussian, binary and mixed data, but zero-inflated models are rarely studied, although they reflect the structure of many data sets in a relevant way. The

objective of this thesis is to infer networks for zero-inflated models.

In this thesis, we will restrict to conditional dependency graphs. The work presented in this thesis is divided into two main parts. The first one concerns graph inference methods based on the estimation of neighbourhoods by a procedure combining ordinal regression models and variable selection methods. The second one focuses on graph inference in a model where the variables are Gaussian zero-inflated by double truncation (right and left).

Keywords: Graph inference, networks, zero-inflated models, regression, Lasso penalisation, variable selection, doubly truncated gaussian data, conditional dependency.