Modèles graphiques non orientés pour des variables aléatoires continues

Auteur : Baum, Carole
Promoteur(s) : Haesbroeck, Gentiane
Faculté : Faculté des Sciences
Diplôme : Master en sciences mathématiques, à finalité spécialisée en statistique
Année académique : 2019-2020
URI/URL : http://hdl.handle.net/2268.2/9252

Université de Liège

Faculté des sciences

Département de mathématiques

Modèles graphiques non orientés pour des variables aléatoires continues

Mémoire de fin d'études présenté en vue de l'obtention du grade de Master en Sciences Mathématiques, à finalité spécialisée en statistique
Promoteur : Gentiane HAESBROECK

Réalisé par

Carole Baum

Année académique 2019-2020

Remerciements

Je tiens à exprimer ma reconnaissance envers toutes les personnes qui m'ont apporté leur aide et soutenue durant la rédaction de ce mémoire ainsi que durant ces années universitaires.

Premièrement, je souhaite remercier ma promotrice Gentiane HAESBROECK de m'avoir encadrée tout au long de ce travail. J'aimerais tout particulièrement souligner sa disponibilité, malgré un emploi du temps fort chargé.

J'aimerai également remercier Madame Stéphanie AERTS et Monsieur Amir ABOUBACAR d'avoir accepté de faire partie des membres de mon jury, ainsi que Madame Émilie CHARLIER, responsable point fixe.

Je remercie également tous mes amis et amies, rencontrés durant ces 5 années d'études, qui m'ont permis de m'épanouir et de profiter de la vie universitaire. J'aimerai aussi exprimer mes amitiés les plus sincères à mes amis de ma région sur lesquels j'ai pu compter.

Je terminerai par remercier ma famille qui m'a soutenue durant les périodes les plus difficiles, en particulier Laurent DOHOGNE ainsi que mes parents et ma soeur pour leur nombreux conseils concernant la rédaction de ce mémoire. Merci à vous tous d'avoir cru en mes capacités et de m'avoir permis de réaliser ces études.

Carole BAUM

Table des matières

Introduction

Un modèle graphique est un graphe qui représente les dépendances entre plusieurs variables aléatoires. Un graphe est composé d'un ensemble de sommets, et d'un ensemble d'arêtes qui joignent certaines paires de sommets. Dans les modèles graphiques, chaque sommet représente une variable aléatoire. Le modèle graphique représentant une fonction de densité multivariée ou un ensemble de données multivariées donne un moyen visuel pour comprendre la distribution jointe de l'ensemble entier des variables aléatoires. Dans un graphe non orienté, les arêtes n'ont pas de flèche directionnelle. Dans ce mémoire, seuls les graphes non orientés, appelés "réseaux de Markov", seront étudiés. Dans ces graphes, l'absence d'arête reliant deux sommets a une signification spéciale : les variables aléatoires correspondantes sont conditionnellement indépendantes, sachant les autres variables. Moins un graphe possède d'arêtes, plus il est facile à interpréter.

Le but de l'étude des modèles graphiques est de pouvoir, à partir d'un ensemble de données, en déduire un modèle graphique adéquat. Pour ce faire, il est nécessaire d'estimer les dépendances entre les variables deux à deux. En effet, les arêtes des graphes sont paramétrées par des valeurs qui encodent la force de la dépendance conditionnelle entre les variables aléatoires aux arêtes correspondantes. Ainsi, une force de dépendance nulle signifie qu'il n'y a pas d'arête dans le graphe.

A partir des données, une estimation de la dépendance entre deux variables ne donnera jamais une valeur exactement nulle. C'est pourquoi, en estimant uniquement ces dépendances, le graphe sera complet et donc inutile dans la compréhension de la distribution jointe des variables. Le problème réside donc dans l'élimination de certaines arêtes dont le potentiel est le plus faible.

Dans ce mémoire, l'étude sera restreinte aux variables continues car les approches d'estimation pour des variables continues ou discrètes sont très différentes. Trois techniques d'estimation de structure de graphe seront présentées avant de les comparer.

Pour commencer, quelques notions utiles sur la théorie des graphes seront développées dans le chapitre 1. Dans ce chapitre sera aussi présentée la base de données qui sera utilisée dans tout ce mémoire pour illustrer les différentes méthodes d'estimation de modèles graphiques.

Dans le deuxième chapitre de ce mémoire, les modèles graphiques gaussiens seront étudiés. Sous la loi normale multivariée, les dépendances entre les variables sont entièrement

connues à travers la matrice de concentration. L'étude de modèles graphiques gaussiens sera donc principalement axée sur l'analyse de cette matrice.

Pour élargir les hypothèses et ne plus se restreindre à des données suivant une loi normale multivariée, le chapitre 3 propose une méthode d'estimation de structure de graphe pour laquelle aucune restriction sur la distribution des données n'est faite. Pour cette technique, chaque variable sera transformée afin que la distribution multivariée obtenue soit une distribution multinormale. La technique du premier chapitre pourra ensuite être appliquée pour obtenir le modèle graphique correspondant.

Les deux premiers chapitres permettent d'obtenir des modèles graphiques pouvant avoir, dans certains cas, un grand nombre d'arêtes. Pour limiter ce nombre d'arêtes, la technique du chapitre 3 propose de restreindre le graphe à une forêt ; c'est à dire que le modèle graphique possèdera au maximum une arête de moins que le nombre de variables. Ainsi, seules les dépendances les plus fortes seront mises en avant à l'aide de cette méthode.

Enfin, le dernier chapitre comparera l'efficacité de ces trois méthodes. A cet effet, le taux d'erreur de chaque méthode sera calculé sur différents exemples. Ce chapitre permettra, par conséquent, de voir quelle méthode il est préférable d'utiliser, sous quelles conditions et dans quel but.

Chapitre 1

Introduction aux graphes de Markov et à la base de données

Dans ce chapitre sont expliquées les notions utiles sur les graphes et leur interprétation dans le domaine statistique, en rapport avec les modèles graphiques. Aucune démonstration ne sera effectuée car ce chapitre permet uniquement de rapeller les bases sur les graphes et de poser les notations. La base de données sera ensuite détaillée. Ce chapitre a été rédigé en faisait référence au chapitre 17 du livre [10] ainsi qu'aux notes de cours de "Théorie des graphes" de M. Rigo.

1.1 Graphes de Markov et leurs propriétés

Définition 1. — Soit V un ensemble et E une partie de $V \times V$. Le *graphe* $\mathcal{G} = (V, E)$ est la donnée du couple (V, E). Les éléments de V sont appelés les sommets de $\mathcal{G}$. Les éléments de E sont appelés les arêtes de $\mathcal{G}$.

— Deux sommets X et Y sont dit *adjacents* s'il y a une arête les joignant, ce qui est noté par (X, Y).

— Un *chemin* $X_1, X_2, \ldots, X_n$ est un ensemble de sommets qui sont reliés, ce qui est noté par $(X_1, \cdots, X_n)$.

— Un graphe *complet* est un graphe pour lequel chaque paire de sommets est reliée par une arête.

— Soient $\mathcal{G} = (V, E)$ et $\mathcal{G}' = (V', E')$ deux graphes. Le graphe $\mathcal{G}'$ est un *sous-graphe* de G si

  ▷ $V' \subseteq V$,

  ▷ $E' \subseteq E \cap (V' \times V')$.

  Ainsi, $\mathcal{G}'$ est un sous-graphe de $\mathcal{G}$ s'il est obtenu en enlevant à $\mathcal{G}$ certains sommets et/ou certaines arêtes.

— Un graphe $\mathcal{G} = (V, E)$ est dit *simple* s'il possède au plus une arête entre 2 sommets et s'il est irréflexif, c'est-à-dire que, quel que soit $v \in V$, (v, v) n'appartient pas à E (i.e., $\mathcal{G}$ ne contient pas de boucle).

— Soit $\mathcal{G} = (V, E)$ un graphe dont les sommets sont ordonnées $V = \{X_1, \cdots, X_n\}$. Sa *matrice d'adjacence* est la matrice $A(\mathcal{G})$ dont l'élément $[A(\mathcal{G})]_{ij}$ est égal au nombre d'arêtes (X_i, X_j) présentes dans le graphe $\mathcal{G}$, $1 \leq i, j \leq n$.

Dans la suite, tous les graphes considérés seront simples, car il n'est pas utile d'étudier la dépendance entre une variable et elle-même. La matrice d'adjacence de ces graphes sera donc uniquement composée de 0 et de 1, et sa diagonale possèdera des valeurs nulles.

Voici quelques exemples de graphes simples. Dans la Figure 1.1(a), (X, Y, Z) forme un chemin mais pas un graphe complet car il manque une arête entre X et Z pour former un graphe complet.

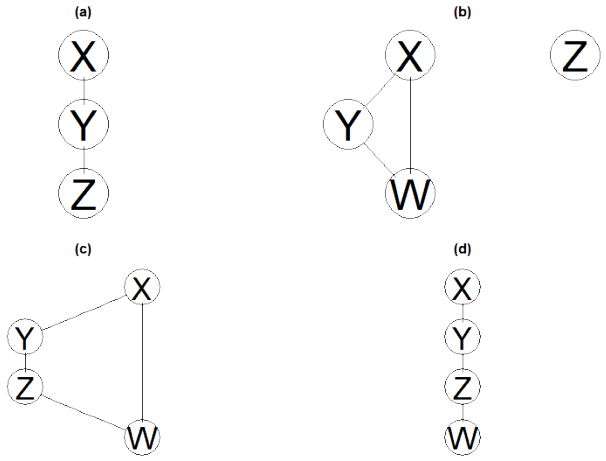

FIGURE 1.1 – Exemples de modèles graphiques non dirigés.

Supposons disposer d'un graphe $\mathcal{G}$ dont l'ensemble des sommets V représente un ensemble de variables aléatoires de distribution jointe p. Dans un graphe de Markov $\mathcal{G}$, l'absence d'une arête entre les sommets X et Y implique que les variables aléatoires correspondantes sont conditionnellement indépendantes sachant les autres variables aux autres sommets ; c'est-à-dire que, sachant la valeur des autres variables du graphe, la probabilité que X ait une certaine valeur reste inchangée, que la valeur de la variable Y soit connue ou non. Cette indépendance est exprimée avec la notation suivante : pas d'arête joignant X et $Y \Leftrightarrow X \amalg Y | \setminus \{X, Y\}$, où "$\setminus\{X, Y\}$" fait référence à tous les autres sommets du graphe. Par exemple dans la Figure 1.1(a), $X \amalg Z|Y$. Ce sont les propriétés d'indépendance par paire de Markov de $\mathcal{G}$.

Définition 2. Si A, B et C sont des sous-ensembles de V, l'ensemble des sommets du graphe, alors C *sépare* A et B si chaque chemin de A vers B passe par un sommet de C.

Par exemple, Y sépare X et Z dans les Figures 1.1(a) et 1.1(d), et Z sépare Y et W dans la Figure 1.1(d). Dans la Figure 1.1(b), Z n'est pas connecté à X, Y et W ; ainsi, les 2 ensembles sont séparés par l'ensemble vide. Dans la Figure 1.1(c), $C = \{X, Z\}$ sépare Y et W.

Les séparateurs ont la propriété de casser le graphe en parties indépendantes conditionnellement. En particulier, dans un graphe de Markov $\mathcal{G}$ avec des sous-ensembles de sommets A, B et C, si C sépare A et B alors $A \amalg B|C$. Ce sont les propriétés globales de Markov de $\mathcal{G}$. Les propriétés globales (indépendance entre 2 ensembles de sommets) et par paire (indépendance entre 2 sommets) de Markov sont équivalentes pour les graphes avec une densité strictement positive. Ces résultats sont utiles pour déduire les relations d'indépendance globale à partir des simples propriétés par paire. Par exemple, dans la Figure 1.1(d), $X \amalg Z|\{Y, W\}$ car c'est un graphe de Markov et qu'aucun lien ne joint X et Z. Mais Y sépare aussi X de Z et W ; c'est pourquoi, à partir des propriétés globales de Markov, $X \amalg Z|Y$ et $X \amalg W|Y$. De manière similaire, $Y \amalg W|Z$.

La propriété globale de Markov permet de décomposer les graphes en parties plus petites et plus faciles à manipuler. Celle-ci amène donc à des simplifications essentielles de calcul et d'interprétation. C'est dans ce but que les graphes sont divisés en cliques.

Définition 3. Une *clique* est un sous-graphe complet ; autrement dit, un ensemble de sommets qui sont tous adjacents aux autres. Une clique est *maximale* si aucun sommet ne peut être ajouté pour qu'elle reste une clique.

Les cliques maximales pour les graphes de la Figure 1.1 sont :

(a) $\{X, Y\}, \{Y, Z\}$

(b) $\{X, Y, W\}, \{Z\}$

(c) $\{X, Y\}, \{Y, Z\}, \{Z, W\}, \{X, W\}$

(d) $\{X, Y\}, \{Y, Z\}, \{Z, W\}$

Certaines méthodes pour estimer la structure de graphes utilisent la notion de clique. En effet, une fonction de densité peut être associée naturellement à un graphe lorsqu'elle peut se factoriser d'une façon bien précise, qui fera intervenir les cliques. Une technique utilisant cette factorisation est développée dans le chapitre 4.

Voici d'autres définitions qui seront utiles pour le chapitre 4.

Définition 4. — Soient $\mathcal{G} = (V, E)$ un graphe et $\mathcal{G}' = (V', E')$ un de ses sous-graphes. On dit que $\mathcal{G}'$ est un *sous-graphe couvrant* $\mathcal{G}$, si $V' = V$ et si

$$\forall v \in V, \exists z \in V : \{z, v\} \in E'.$$

Autrement dit, tout sommet de $\mathcal{G}$ est une extrémité d'une arête de E'.

— Un graphe simple, non orienté $\mathcal{G} = (V, E)$ est un *arbre* s'il est connexe et sans cycle. Une *forêt* est un graphe simple, non orienté dont chaque composant connexe est un arbre.

Dans ce chapitre, l'intérêt sera plus particulièrement porté sur les sous-graphes couvrants qui sont des arbres, ceux-si seront alors naturellement nommés sous-arbres couvrants ou arbres couvrants.

1.2 Présentation de la base de données socio-économiques de la Wallonie

La base de données choisie permet d'analyser les dépendances entre différentes caractéristiques économiques des communes wallonnes. Cette base de données a été créée à partir des chiffres repris sur le site `https://walstat.iweps.be/walstat-accueil.php`, consulté en mars 2019.

Celle-ci contient 262 observations, c'est à dire les 262 communes wallonnes, et 19 variables. Ces variables sont toutes continues et ne possèdent pas de valeurs manquantes. Voici la description des différentes variables socio-économiques :

TMortH Taux de mortalité standardisé sur l'âge pour les hommes (2005-2014). Le taux standardisé par âge et par sexe est le taux que l'on observerait dans la population étudiée si elle avait la même structure d'âge qu'une population de référence (ici la population européenne de 2013). Il est calculé en pondérant les taux de mortalité par âge observé dans la sous-population par la structure d'âge de la population de référence.

TMortF Taux de mortalité standardisé sur l'âge pour les femmes (2005-2014).

ITrafic Intensité du trafic routier : millions de véhicules par km (2005). Un véhicule par km représente le déplacement d'un véhicule sur la distance d'un km sur le réseau routier de l'entité.

ImmVe Immatriculation de véhicules neufs (2018). Cette statistique donne un aperçu de tous les véhicules motorisés neufs immatriculés en Belgique auprès du service d'immatriculation de véhicules du SPF Mobilité pour une période donnée (ici l'année 2018). Types de véhicules concernés : voitures de tourisme, autobus et autocars, camions, tracteurs routiers, tracteurs agricoles, véhicules spéciaux et motocycles. Types de carburant : essence, diesel, gaz ou électricité.

TEmplH Taux d'emploi administratif des hommes (2016). L'indicateur rapporte le nombre de personnes de sexe masculin qui ont effectivement un emploi (population active occupée) à la population des hommes, en moyenne annuelle.

TemplF Taux d'emploi administratif des femmes (2016). L'indicateur rapporte le nombre de personnes de sexe féminin qui ont effectivement un emploi (population active occupée) à la population des femmes, en moyenne annuelle.

TchH Taux de chômage administratif des hommes (2016). L'indicateur rapporte le nombre de personnes de sexe masculin au chômage à la population des hommes, en moyenne annuelle.

TChF Taux de chômage administratif des femmes (2016). L'indicateur rapporte le nombre de personnes de sexe féminin au chômage à la population des femmes, en moyenne annuelle.

QOrd Quantité d'ordures ménagères brutes collectée par habitant (2016). L'indicateur donne la quantité collectée d'ordures ménagères brutes (OMB) des ménages, exprimée en kg, par habitant et par an. Les OMB correspondent au contenu des poubelles "tout venant" non triées, en ce compris les quantités d'OMB assimilées (OMB des commerces, écoles, voiries, marchés, etc., collectées en même temps que les OMB des ménages par les communes ou les intercommunales).

CNrj Consommation d'énergie totale en GWh (2015).

Elec Production d'électricité en GWh (2015).

TPartElec Taux de participation aux élections communales (2018). Ce taux se calcule en faisant le rapport entre le nombre de bulletins déposés et le nombre d'électeurs inscrits sur les listes électorales. Le vote étant obligatoire en Belgique, ce taux est naturellement élevé. Cependant, bien que les citoyens belges s'exposent théoriquement à une sanction en n'allant pas voter, une partie d'entre eux ne remplit pas ce devoir.

PopTot Population totale : nombre d'habitants dans l'entité (2018).

TNat Taux brut de natalité (2017). Rapport (exprimé en ‰ habitants) du nombre de naissances vivantes de l'année à la population totale moyenne de l'année (somme, divisée par 2, de la population au 1er janvier et de celle au 31 décembre de l'année).

AgeMoy Age moyen de la population (2018). Il se calcule en faisant la somme des années vécues par la population divisée par la population totale de l'entité à l'année de référence. Par exemple, une personne âgée de 42 ans l'année de référence "compte" pour 42 années vécues.

SolMigr Solde migratoire total (2017). Le solde migratoire total est la différence entre la population du 1er janvier et celle du 31 décembre moins le solde naturel (les naissances moins les décès) ; rapportée à la population totale moyenne de l'année (somme, divisée par 2, de la population au 1er janvier et de celle au 31 décembre de l'année).

KmRoute Kilomètres de réseau routier revêtu total (2005). Pour calculer la longueur du réseau routier revêtu, on additionne les kilomètres des différents réseaux. Sont considérés : les autoroutes (AR) et les rings à statut autoroutier, les routes nationales (RN) (non compris les entrées/sorties d'AR et RN, parkings...), les routes régionales (y compris les entrées/sorties d'AR et de RN, parkings...), toutes les routes provinciales, les routes communales carrossables, à l'exclusion, en principe, des chemins agricoles et sentiers et des voiries de grande circulation comprises dans le réseau communal.

NbAcc Nombre total d'accidents de la circulation (2017). Sont considérés comme "accidents", les accidents de la circulation routière impliquant au moins un véhicule qui occasionne des dommages corporels et qui se sont produits sur la voie publique. Un accident entre plus de deux véhicules est considéré comme un seul accident. Les accidents sont comptabilisés dans l'entité où ils ont eu lieu.

RevMoy Revenu moyen par habitant en euro (2016). Les statistiques fiscales sont établies sur la base des déclarations à l'impôt des personnes physiques au lieu de résidence.

Ces différentes variables ont été choisies pour différentes raisons. Tout d'abord, il est possible de penser, à priori, que certaines variables sont dépendantes. Par exemple, les taux d'emploi et de chômage pour les deux sexes sont intuitivement fortement liés. De plus, les variables concernant le trafic routier (comme l'intensité du trafic routier, le nombre de véhicules neufs immatriculés, le nombre de kilomètres de réseau routier revêtu, le nombre total d'accidents de la circulation) risquent d'avoir, elles aussi, une forte dépendance. Ensuite, d'autres variables ont été choisies en fonction de l'actualité ; c'est-à-dire de l'enjeu environnemental. Il est alors question des variables quantité d'ordures ménagères, consommation et production d'électricité ainsi que des variables concernant le trafic routier. Les autres variables sont des critères économiques intéressants à comparer et à étudier.

Chapitre 2

Modèles graphiques gaussiens

Une première étape dans l'étude des modèles graphiques consiste à étudier les vecteurs aléatoires multinormaux. En effet, pour ceux-ci, les dépendances entre les variables sont reprises dans la matrice de corrélation. Ce chapitre permet de construire le modèle graphique représentant les données à partir de l'estimation de la matrice de covariances.

Lorsque la matrice de covariances est estimée, aucune entrée de celle-ci ne sera exactement nulle. C'est pourquoi un paramètre de pénalisation sera introduit pour forcer certaines valeurs à être nulles. Une technique de pénalisation est développée dans ce chapitre en faisant référence au chapitre 17 du livre [10].

2.1 Propriétés de la loi multinormale

L'hypothèse faite sur les données dans ce chapitre est que les observations suivent une distribution gaussienne multivariée de dimension d, de moyenne μ et de matrice de covariance Σ. Cette matrice est supposée définie positive, ainsi la base de données est supposée non dégénérée. De plus, le nombre n d'observations doit être suffisamment grand par rapport à d. Une règle souvent utilisée est la condition $\frac{n}{d} > 5$, mentionnée notamment dans [11]. La particularité du cas gaussien est que l'indépendance entre deux variables est équivalente à une corrélation ou une covariance nulle entre ces mêmes variables. C'est pourquoi la matrice de covariances des variables sera étudiée.

Une des propriétés de la distribution gaussienne multivariée est que toutes les distributions conditionnelles sont aussi gaussiennes. Supposons avoir des variables $X_1, \cdots, X_d$ dont la matrice de covariances est inversible. L'inverse de cette matrice, la matrice de concentration $\Theta = \Sigma^{-1}$, contient les informations sur la covariance partielle entre les variables, c'est-à-dire, la covariance entre la paire X_i et X_j, conditionnellement à toutes les autres variables. En particulier, si la composante ij de Θ est égale zéro, alors les variables X_i et X_j sont conditionnellement indépendantes, sachant les autres variables.

Pour s'en convaincre, la distribution conditionnelle d'une variable par rapport au reste (où le rôle de Θ est explicite) est analysée. La proposition suivante a été rédigée en référence à la proposition 3.13 de [6].

Proposition 1. *Soit $X \sim \mathcal{N}_d(\mu, \Sigma)$, avec $\Sigma \in \mathbb{R}^{d\times d}$ définie positive et $\mu \in \mathbb{R}^d$. Si X est partitionné par $X = (Z, Y)$ où $Z = (X_1, \ldots, X_{d-1})$ consiste en les $d-1$ premières variables et $Y = X_d$ la dernière, alors, la distribution conditionnelle de Y sachant Z est donnée par :*

$$Y|Z = z \sim \mathcal{N}(\mu_Y + (z - \mu_Z)^T \Sigma_{ZZ}^{-1} \sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}), \tag{2.1}$$

où Σ est partitionné par

$$\Sigma = \begin{pmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix} \text{ et } \mu \text{ par } \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}. \tag{2.2}$$

Démonstration. Il faut tout d'abord remarquer que Σ_{ZZ} est défini positif puisque c'est le cas de Σ et donc qu'il est inversible. En effet, en procédant par l'absurde, il est facile de s'en convaincre. Soit x un vecteur. Si $x^T \Sigma_{ZZ} x \leq 0$ alors,

$$\begin{pmatrix} x & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} x^T \Sigma_{ZZ} & x^T \sigma_{ZY} \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = x^T \Sigma_{ZZ} x \leq 0.$$

Ce qui est impossible car pour tout y, donc en particulier pour $y = (x, 0)$, $y^T \Sigma y > 0$.

Une deuxième remarque à faire est que les variables $Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z$ et Z sont indépendantes. En effet, leur densité jointe est la loi normale, par définition des vecteurs gaussiens, et la covariance entre ces deux variables est donnée par

$$\begin{aligned} \text{cov}(Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z, Z) &= \text{cov}(Y, Z) - \text{cov}(\sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z, Z) \\ &= \sigma_{ZY}^T - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \Sigma_{ZZ} = 0 \end{aligned}$$

Ainsi ces variables sont non corrélées, donc indépendantes.

De plus, la variance de $Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z$ est donnée par $\sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}$. En effet,

$$\begin{aligned} \text{var}(Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z) &= \text{var}(Y) + \text{var}(\sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z) - 2\text{cov}(Y, \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z) \\ &= \sigma_{YY} + \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \Sigma_{ZZ} \Sigma_{ZZ}^{-1} \sigma_{ZY} - 2\sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY} \\ &= \sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}. \end{aligned}$$

Pour trouver la loi de la variable $Y|Z = z$, la fonction caractéristique de cette variable est calculée ; ce qui mène à

$$\begin{aligned} \mathbb{E}[e^{itY}|Z = z] &= \mathbb{E}[e^{it[Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z + \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z]}|Z = z] \\ &= e^{it[\sigma_{ZY}^T \Sigma_{ZZ}^{-1} z]}\, \mathbb{E}[e^{it[Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z]}|Z = z] \\ &= e^{it[\sigma_{ZY}^T \Sigma_{ZZ}^{-1} z]}\, \mathbb{E}[e^{it[Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z]}], \end{aligned}$$

où l'indépendance entre Z et $Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z$ a été utilisée à la dernière ligne.

Or, il est clair que

$$Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} Z \sim \mathcal{N}(\mu_Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \mu_Z, \sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}),$$

au vu de la deuxième remarque.

Ainsi,

$$\begin{aligned}\mathbb{E}[e^{itY}|Z=z] &= e^{it[\sigma_{ZY}^T \Sigma_{ZZ}^{-1} z]} e^{it[\mu_Y - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \mu_Z]} e^{-\frac{t^2}{2}[\sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}]} \\ &= e^{it[\mu_Y + \sigma_{ZY}^T \Sigma_{ZZ}^{-1}(z-\mu_z)] - \frac{t^2}{2}[\sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY}]}.\end{aligned}$$

La conclusion en découle, car la loi normale est entièrement déterminée par sa fonction caractéristique. □

La moyenne conditionnelle dans (2.1) a exactement la même forme que la régression linéaire en population multiple de Y par rapport à Z, avec le coefficient de régression $\beta = \Sigma_{ZZ}^{-1} \sigma_{ZY}$. Si Θ est partitionné de la même façon, comme $\Sigma\Theta = I_d$, la formule pour partitionner l'inverse donne

$$\theta_{ZY} = -\theta_{YY} \Sigma_{ZZ}^{-1} \sigma_{ZY}, \tag{2.3}$$

où $1/\theta_{YY} = \sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1} \sigma_{ZY} > 0$. C'est pourquoi

$$\beta = \Sigma_{ZZ}^{-1} \sigma_{ZY} \tag{2.4}$$

$$= -\theta_{ZY}/\theta_{YY}. \tag{2.5}$$

Deux choses peuvent être déduites de la forme de la distribution conditionnelle normale :

— La dépendance de Y sur Z dans (2.1) est dans le terme de la moyenne seulement. Une composante nulle dans β et donc une valeur nulle de θ_{ZY} signifie que l'élément correspondant de Z est conditionnellement indépendant de Y, sachant le reste.

— Cette dépendance peut être étudiée à l'aide d'une régression linéaire multiple.

Ainsi, Θ capture toute l'information de second ordre requise pour décrire la distribution conditionnelle de chaque sommet, sachant les autres.

2.2 Estimation des paramètres du graphe

Le but de cette section est d'estimer les paramètres d'un graphe non dirigé qui approxime la distribution jointe d'une réalisation de X.

Avant d'estimer les paramètres du graphe, il est judicieux de standardiser les données. En effet, la structure du graphe dépend de la matrice de covariances qui sera différente si les données sont standardisées ou non. Si les variables constituant les données sont mesurées selon une unité ou une échelle différente, alors il sera préférable de standardiser les données.

2.2.1 Graphe complet

Une première étape consiste à créer un graphe complet. Soient n réalisations normales multivariées x_i, $i = 1, \ldots, n$ de moyenne μ et de matrice de covariances Σ ; avec $\Theta = \Sigma^{-1}$. Soit

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{2.6}$$

la matrice de covariances empiriques, avec $\bar{x}$ le vecteur des moyennes empiriques.

Soit $X \sim \mathcal{N}_d(\mu, \Sigma)$, avec $\Sigma \in \mathbb{R}^{d \times d}$ définie positive et $\mu \in \mathbb{R}^d$. La log-vraisemblance de cette variable multinormale s'obtient facilement à l'aide des propriétés de la trace. En ignorant les constantes et à un facteur multiplicatif près, celle-ci peut être écrite comme

$$l(\mu, \Theta; x_1, \ldots, x_n) = \ln \det \Theta - \operatorname{tr}(\Theta S) - (\bar{x} - \mu)^T \Theta (\bar{x} - \mu). \tag{2.7}$$

La maximisation de la vraisemblance pour μ se fait indépendamment de la valeur de Θ, il est donc possible d'évacuer cette estimation. Pour maximiser $l(\mu, \Theta; x_1, \ldots, x_n)$ par rapport à μ, puisque μ n'intervient que dans un terme et que Θ est définie positive, il faut que $\hat{\mu} = \bar{x}$. Ainsi la vraisemblance déjà maximisée pour μ devient

$$l(\Theta; x_1, \ldots, x_n) = \ln \det \Theta - \operatorname{tr}(S\Theta). \tag{2.8}$$

L'estimateur du maximum de vraisemblance pour Θ est donc S^{-1}, en utilisant le fait que la dérivée de $\ln \det \Theta$ est égale à Θ^{-1} (résultat obtenu à la page 641 du livre [4]). Pour que la matrice de covariances S soit inversible, il est nécessaire que $n > d$, sinon la matrice S est singulière.

Si le graphe est créé à partir de la matrice S^{-1}, alors celui-ci sera complet. En effet, aucune composante de cette matrice ne sera nulle, donc aucune arête ne sera retirée du graphe. En pratique, cette construction n'apporte aucune information exploitable sur les dépendances entre les variables. C'est pourquoi il va falloir contraindre certaines composantes de la matrice de concentration estimée à être nulles.

2.2.2 Graphe de structure connue

Dans cette section, la structure du graphe est connue, donc, en particulier, l'ensemble E des arêtes du graphe est connu. Si certaines arêtes sont manquantes, par les propriétés de la distribution gaussienne, cela implique que les entrées correspondantes de $\Theta = \Sigma^{-1}$ sont nulles. C'est pourquoi il faut maintenant maximiser (2.8) sous les contraintes que quelques ensembles de paramètres prédéfinis sont nuls. C'est un problème d'optimisation convexe avec des contraintes d'égalité.

Plusieurs méthodes existent pour résoudre ce problème d'optimisation. Dans ce mémoire, un approche basée sur la régression linéaire sera suivie en référence au chapitre 17

du livre [10]. Une autre méthode qui est souvent utilisée est la méthode IPS (Iterative Proportional Scaling) dévelopée dans [25].

Pour contraindre le maximum de vraisemblance, une constante de Lagrange est ajoutée pour toutes les arêtes manquantes

$$l_C(\Theta; x_1, \ldots, x_n) = \ln\det\Theta - \mathrm{tr}(S\Theta) - \sum_{(j,k)\notin E} \gamma_{jk}\theta_{jk}. \tag{2.9}$$

L'équation de gradient pour maximiser (2.9) peut être écrite sous la forme

$$\Theta^{-1} - S - \Gamma = 0. \tag{2.10}$$

La matrice Γ est la matrice des paramètres de Lagrange avec des valeurs différentes de zéro pour toutes les paires dont l'arête est absente. Comme le graphe ne peut pas posséder de boucles, les arêtes du type (i, i) ne font jamais partie du graphe. Celles-ci ne feront donc pas l'objet de l'étude des arêtes manquantes et la diagonale de Γ sera toujours nulle.

La régression est utilisée pour estimer Θ et son inverse, noté $W = \hat{\Theta}^{-1} = \hat{\Sigma}$ en utilisant (2.10). Cette régression sera effectuée sur une ligne et une colonne à la fois. Pour simplifier les notations, la dernière ligne et la dernière colonne sont considérées. Le bloc du haut à droite de l'équation (2.10) peut être écrit comme suit :

$$w_{12} - s_{12} - \gamma_{12} = 0. \tag{2.11}$$

Les matrices sont partitionnées en deux parties : la première partie contenant les $d-1$ premières lignes et colonnes et la seconde partie les d-ème ligne et colonne. Avec W et son inverse $\hat{\Theta}$ partitionnés de manière similaire, cela donne

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \hat{\Theta}_{11} & \hat{\theta}_{12} \\ \hat{\theta}_{12}^T & \hat{\theta}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{d-1} & 0 \\ 0^T & 1 \end{pmatrix}. \tag{2.12}$$

Ce qui implique

$$w_{12} = -W_{11}\hat{\theta}_{12}/\hat{\theta}_{22} = W_{11}\beta, \tag{2.13}$$

où $\beta = -\hat{\theta}_{12}/\hat{\theta}_{22}$ comme dans (2.5). Maintenant, en remplaçant (2.13) dans (2.11), l'équation suivante est obtenue

$$W_{11}\beta - s_{12} - \gamma_{12} = 0. \tag{2.14}$$

Cela peut être interprété comme les $d-1$ équations à estimer pour la régression contrainte de X_d par rapport aux autres variables, sauf que la matrice S_{11} des covariances empiriques est remplacée par W_{11}, la matrice de covariances pour le modèle estimé actuellement.

Maintenant, l'équation (2.14) peut être résolue par une simple régression par sous-ensembles. Si $d-q$ éléments sont non nuls dans γ_{12}, i.e., $d-q$ arêtes contraintes à être nulles ; alors, ces $d-q$ lignes n'apportent pas d'information et peuvent être retirées. De

plus, β peut être réduit à β^* en enlevant ces $d-q$ éléments nuls ; pour obtenir ainsi le système d'équation

$$W_{11}^*\beta^* - s_{12}^* = 0 \tag{2.15}$$

qui a pour solution $\beta^* = {W_{11}^*}^{-1}s_{12}^*$. Il reste à ajouter les $d-q$ zéros pour obtenir $\hat{\beta}$.

Il est facile de montrer que

$$\frac{1}{\hat{\theta}_{22}} = w_{22} - w_{12}^T\beta, \tag{2.16}$$

en utilisant la formule de l'inverse partitionné. De plus, $w_{22} = s_{22}$, comme la diagonale de Γ est nulle dans (2.10).

Cela conduit à la simple procédure itérative donnée dans l'algorithme 1 pour estimer W et son inverse $\hat{\Theta}$, sujets aux contraintes des arêtes manquantes.

Algorithme 1. *Un algorithme de régression modifié pour estimer un modèle graphique gaussien avec une structure connue.*

1. *Initialiser* $W = S$
2. *Répéter pour* $j = 1, 2, \ldots, d, 1, \ldots$ *jusqu'à la convergence :*
   (a) *Partitionner la matrice* W *en 2 parties. En changeant éventuellement l'ordre des lignes et des colonnes, la première partie contiendra tout sauf les j-ème ligne et colonne ; et la deuxième, les j-ème ligne et colonne.*
   (b) *Résoudre* $W_{11}^*\beta^* - s_{12}^* = 0$ *pour les paramètres* β^* *des arêtes non contraintes, en utilisant le système réduit d'équations comme dans* (2.15)*. Obtenir* $\hat{\beta}$ *en ajoutant à* $\hat{\beta}^*$ *les zéros aux positions appropriées.*
   (c) *Mettre à jour* $w_{12} = W_{11}\hat{\beta}$
3. *Dans le cycle final (pour chaque j) résoudre* $\hat{\theta}_{12} = -\hat{\beta}\hat{\theta}_{22}$*, avec* $\dfrac{1}{\hat{\theta}_{22}} = s_{22} - w_{12}^T\hat{\beta}$

2.2.3 Estimation de la structure du graphe

Dans la plupart des cas, la place des arêtes à enlever au graphe est inconnue, c'est pourquoi il est utile de pouvoir déduire ces places à partir des données elles-mêmes. Pour ce faire la méthode lasso est utilisée et le développement sera similaire à celui fait précédemment. La méthode lasso consiste à maximiser la log-vraisemblance pénalisée

$$\log \det\Theta - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1, \tag{2.17}$$

où $||\Theta||_1$ est la norme L_1 (la somme des valeurs absolues des éléments de Σ^{-1}), et où les constantes sont ignorées. Remarquons que la plupart des auteurs prennent en compte la diagonale de la matrice Θ pour calculer la norme, sauf quelques-uns, comme *Yuan et Liu* dans [30]. Dans ce mémoire, la diagonale fera partie de la norme car elle pénalisera

d'avantage la log-vraisemblance. Il faut remarquer que celle-ci possède des valeurs qui ont des grandeurs similaires car les données ont été préalablement standardisées.

Cette vraisemblance pénalisée est une fonction convexe de Θ. Cela signifie qu'il est possible d'adapter le lasso pour obtenir le paramètre permettant de maximiser la log-vraisemblance pénalisée. En pratique, il suffit simplement de remplacer l'étape (b) de régression modifiée dans l'algorithme 1 par une étape de lasso modifiée. Cette procédure est détaillée dans les paragraphes suivants.

L'analogue de l'équation de gradient est maintenant

$$\Theta^{-1} - S - \lambda \text{Sign}(\Theta) = 0, \tag{2.18}$$

avec $\text{Sign}(\theta_{jk}) = \text{sign}(\theta_{jk})$ si $\theta_{jk} \neq 0$, et $\text{Sign}(\theta_{jk}) \in [-1, 1]$ si $\theta_{jk} = 0$. En effectuant le même développement que dans la section précédente, l'analogue de (2.14) est atteint

$$W_{11}\beta - s_{12} + \lambda \text{Sign}(\beta) = 0. \tag{2.19}$$

(β et θ_{12} ont des signes opposés). Il reste à montrer que ce système est équivalent à l'estimation des équations pour la régression lasso.

Soient $\mathbf{y}$ la variable de résultat d'une régression linéaire et $\mathbf{Z}$ la matrice des variables explicatives. Ainsi la méthode de régression lasso consiste à minimiser, à un facteur multiplicatif près,

$$\frac{1}{2}(\mathbf{y} - \mathbf{Z}\beta)^T(\mathbf{y} - \mathbf{Z}\beta) + \lambda||\beta||_1. \tag{2.20}$$

Le gradient de cette expression est donné par

$$\mathbf{Z}^T\mathbf{Z}\beta - \mathbf{Z}^T\mathbf{y} + \lambda \text{Sign}(\beta) = 0. \tag{2.21}$$

Ainsi, $\mathbf{Z}^T\mathbf{y}$ est l'analogue de s_{12} et $\mathbf{Z}^T\mathbf{Z}$ est remplacé par W_{11}, l'estimation de la matrice de covariance du modèle actuel.

Cette procédure est appelée *graphical lasso* et est résumée dans l'algorithme 2.

Algorithme 2. *Graphical lasso.*

1. *Initialiser $W = S + \lambda \boldsymbol{I}$. La diagonale de W reste inchangée dans ce qui suit.*
2. *Répéter pour $j = 1, 2, \ldots, d, 1, \ldots$ jusqu'à la convergence :*
   (a) *Partitionner la matrice W en 2 parties. La première contiendra tout sauf les j-ème ligne et colonne ; et la deuxième, les j-ème ligne et colonne.*
   (b) *Résoudre $W_{11}\beta - s_{12} + \lambda Sign(\beta) = 0$, en utilisant l'algorithme de descente de coordonnées cycliques* (2.22) *pour le lasso modifié.*
   (c) *Mettre à jour $w_{12} = W_{11}\hat{\beta}$*
3. *Dans le cycle final (pour chaque j), résoudre $\hat{\theta}_{12} = -\hat{\beta}\hat{\theta}_{22}$, avec $\dfrac{1}{\hat{\theta}_{22}} = s_{22} - w_{12}^T\hat{\beta}$*

Voici quelques explications pour la méthode de descente de coordonnées pour l'algorithme *graphical lasso*. Cette technique est détaillée dans [7]. En posant $\mathbf{V} = \mathbf{W}_{11}$, la mise à jour a la forme

$$\beta_j \leftarrow S(s_{12j} - \sum_{k \neq j} V_{kj}\hat{\beta}_k, \lambda)/V_{jj} \tag{2.22}$$

pour $j = 1, 2, \ldots, d-1, 1, 2, \ldots, d-1, \ldots$, où S est l'opérateur à seuil doux

$$S(x, t) = \text{sign}(x)(|x| - t)_+, \tag{2.23}$$

et où s_{12j} désigne la $j^{\text{ème}}$ composante du vecteur s_{12}. La procédure effectue un cycle parmi les variables estimées jusqu'à la convergence.

Les éléments de la diagonale w_{jj} de la matrice des solutions $\mathbf{W}$ sont simplement $s_{jj} + \lambda$. En effet, ils sont fixés à l'étape 1 de l'algorithme 2 et la modification des éléments de $\mathbf{W}$ a lieu à l'étape 2(c). De plus, seuls les éléments de w_{12}, qui ne se situent pas sur la diagonale, sont mis à jour.

Dans ce mémoire sera utilisé le logiciel R pour effectuer tous les graphiques et calculs éventuels. Dans celui-ci sont implémentés plusieurs algorithmes permettant de construire des graphes basés sur une distribution normale multivariée. Pour construire des modèles graphiques basés sur la théorie développée dans ce chapitre, la fonction *glasso* de la librairie du même nom sera utilisée pour estimer la structure du graphe. Celle-ci a été implémentée en référence à l'article [8] et est basée sur la théorie développée dans ce chapitre et, notamment, sur l'algorithme 2. Pour cette fonction, une valeur du paramètre de pénalisation doit être donnée en entrée ainsi que la matrice de covariances des données. D'autres fonctions, telles que *huge* (qui permet aussi d'effectuer l'algorithme *graphical lasso*) et *shock* (qui estime la matrice de covariances par une matrice par blocs), peuvent, néanmoins, être utilisées pour la construction de modèles graphiques. La librairie *SIN* permet aussi de construire des modèles graphiques gaussiens à partir d'une autre méthode.

Afin de trouver la meilleure valeur pour le paramètre λ de la pénalisation lasso, plusieurs critères peuvent être utilisés. Dans ce mémoire, la minimisation des critères AIC et BIC sera utilisée. En toute généralité, ces mesures sont définies par

$$\text{AIC } = -2l(\mu, \Theta; x_1, ..., x_n) + k.2$$

et

$$\text{BIC } = -2l(\mu, \Theta; x_1, ..., x_n) + k \ln n$$

où k est le nombre de paramètres à estimer et où

$$l(\mu, \Theta; x_1, ..., x_n) = -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln \det\Theta - \frac{n}{2}\text{tr}(S\Theta)$$

(critères développés dans [15]). Dans le contexte d'estimation de la covariance, le nombre de paramètres à estimer peut être très élevé. L'article [30] suggère de modifier les mesures comme suit :

$$\text{AIC } = -2l(\mu, \Theta; x_1, ..., x_n) + \Big(d + \sum_{i \leq j} \hat{e}_{ij}(\lambda)\Big).2$$

et

$$\text{BIC } = -2l(\mu, \Theta; x_1, ..., x_n) + \Big(d + \sum_{i \leq j} \hat{e}_{ij}(\lambda)\Big) \ln n$$

où $\hat{e}_{ij} = 0$ si $\hat{\Theta}_{ij} = 0$ et $\hat{e}_{ij} = 1$ sinon.

Plus la valeur de λ est faible, plus le graphe possédera d'arêtes. Une valeur trop élevée pour λ pourrait manquer certaines relations entre les variables ; alors qu'une valeur trop faible rend le graphe inexploitable car il possédera trop d'arêtes, voire toutes. Ces deux critères sont implémentés dans R à la main.

2.3 Exemples de modèles graphiques

Avant d'analyser le graphe obtenu pour les caractéristiques des communes wallonnes, il est intéressant de tester la technique d'estimation *glasso* sur des données numériques. En effet, pour ces dernières, les dépendances entre les variables sont connues à priori. Cela permettra donc de comparer la structure du graphe obtenue avec les dépendances attendues à priori. Après ces quelques exemples numériques, les graphes créés à partir de la base de données socio-économiques de la Wallonie seront représentés.

2.3.1 Exemples numériques sous la normalité

Pour commencer, voici quelques exemples de modèles graphiques sur des données simulées. Cela permet de montrer que la méthode utilisée pour estimer les graphes est efficace. Pour chaque exemple, $n = 1000$ données ont été simulées. Ainsi, il est clair que n est suffisamment grand par rapport au nombre de variables qui est de $d = 10$.

Plusieurs exemples sont réalisés de manière à faire varier la structure de la matrice de covariances des données. Le vecteur moyen sera toujours le vecteur nul, sans perte de généralités, car celui-ci n'impacte pas l'estimation de la structure du graphe. Les éléments de la diagonale de la matrice de covariances seront toujours égaux à 1, au vu de la standardisation nécessaire pour construire le graphe.

Cas indépendant

Un premier exemple consiste à générer des données selon une loi normale multivariée de matrice de covariances $\Sigma = I$. Comme les variables sont indépendantes, le graphe ne devrait posséder aucune arête, même pour une valeur assez faible du paramètre de pénalisation ; ce que confirme la Figure 2.1 représentant le modèle graphique d'une normale multivariée standard tel qu'estimé à l'aide de la procédure *glasso* pour un paramètre de pénalisation arbitraire $\lambda = 0.1$.

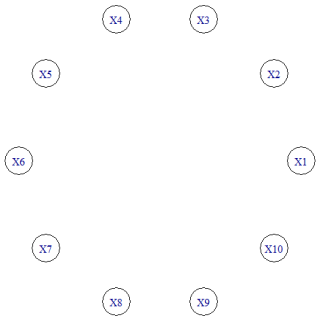

FIGURE 2.1 – Modèle graphique pour des variables qui suivent une loi normale multivariée standard, avec $\lambda = 0.1$.

La Figure 2.2 représente la variation du nombre d'arêtes en fonction de λ, tout en mettant en évidence la valeur sélectionnée par les critères AIC et BIC. Au vu de cette figure, il est possible de voir que le nombre d'arêtes conservées, lorsque les critères AIC et BIC sont utilisés pour trouver la valeur du paramètre de pénalisation, sera strictement positif. Ainsi certaines arêtes seront conservées à tort.

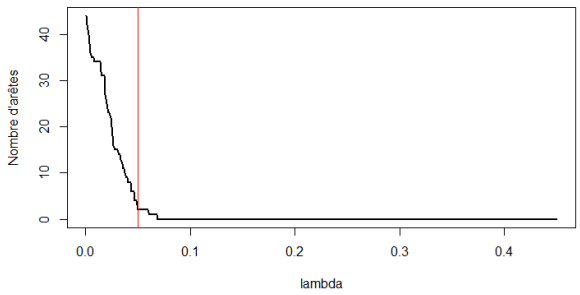

FIGURE 2.2 – Variation du nombre d'arêtes du graphe en fonction de λ, pour des variables qui suivent une loi normale multivariée standard. La ligne rouge représente la valeur de λ qui minimise les critères AIC et BIC.

Cas équi-corrélé

Pour ce deuxième exemple, des données sont générées selon une loi normale multivariée de matrice de covariances équi-corrélée. Dans tout ce mémoire, la notation suivante sera utilisée pour décrire une matrice équi-corrélée de dimension d et dont la corrélation est égale à ρ :

$$\Sigma_{\rho,d} = \rho(\mathbf{1}\mathbf{1}^T - I_d) = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix},$$

où $\mathbf{1}$ est le vecteur de taille d dont toutes les composantes sont égales à 1.

Comme tous les couples de variables ont la même corrélation, les corrélations estimées devraient être toutes presque identiques. Ainsi, l'algorithme du lasso devrait soit laisser toutes les arêtes, soit toutes les supprimer ; ce qui est illustré par la Figure 2.3 lorsque la matrice de covariances est $\Sigma_{0.2,10}$. A gauche de cette figure se trouve le graphe créé avec la pénalisation $\lambda = 0.1$, tandis qu'elle est de 0.3 pour celui de droite.

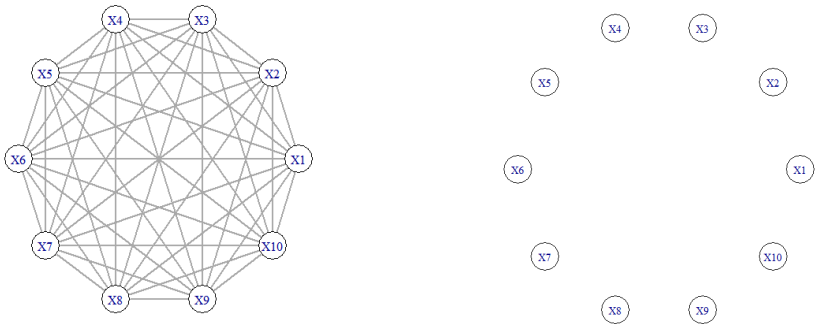

FIGURE 2.3 – Modèle graphique pour des variables équi-corrélées avec $\rho = 0.2$. A gauche se trouve le graphe créé avec la pénalisation $\lambda = 0.1$, tandis qu'elle est de 0.3 pour celui de droite.

Le paramètre de pénalisation λ va permettre de modifier le nombre d'arêtes présentes dans le graphe. Une petite valeur pour le paramètre éliminera peu d'arêtes. A partir d'un certain seuil pour λ, un grand nombre d'arêtes sera supprimé, c'est pourquoi dans la Figure 2.3, le graphe de droite ne possède plus d'arête. La figure 2.4 permet aussi de voir la variation du nombre d'arêtes présentes dans le graphe en fonction de λ. Ce nombre d'arêtes dépendra aussi de la corrélation entre les variables. Plus ρ est grand, plus λ devra être grand pour observer un graphe sans arêtes. Le passage du graphe complet au graphe sans arête

s'effectue de façon assez verticale au niveau de la valeur de ρ. De plus, cette verticalité est de plus en plus marquée lorsque ρ augmente. Le cas où $\rho = 1$ n'est pas considéré car le cas de données dégénérées a été exclu. Des valeurs négatives pour ρ peuvent également être choisies en gardant en tête que la matrice de covariances doit rester définie positive. Pour que cette condition soit vérifiée, il faut que $\rho > \frac{-1}{d-1}$ ce qui est égal à -0.1111 pour $d = 10$. Lorsque $\rho = -0.1$, la courbe représentant le nombre d'arêtes du graphe est similaire à la courbe obtenue pour $\rho = 0.1$. Elle n'est pas représentée sur le graphe de la Figure 2.4 par soucis de clarté du graphe.

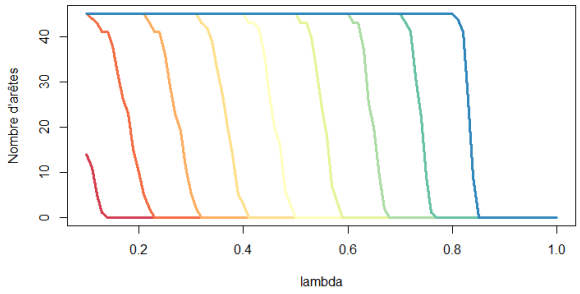

FIGURE 2.4 – Variation du nombre d'arêtes dans le graphe en fonction de la variation de λ et de ρ, pour des variables équi-corrélées. Le paramètre de corrélation ρ varie entre 0.1 et 0.9 par pas de 0.1 et les différentes courbes sont dans l'ordre démarrant de la gauche vers la droite.

Cas auto-corrélé

L'étape suivante consiste à supposer que les variables sont liées par un taux de corrélation dépendant de la place des variables. Cette structure de corrélation est typiquement rencontrée en séries temporelles ou en statistique spatiale. Pour obtenir cette structure de corrélation, des données sont générées selon une loi normale multivariée de matrice de covariances de la forme

$$\begin{pmatrix} 1 & \rho^{|1-2|} & \dots & \rho^{|1-d|} \\ \rho^{|2-1|} & 1 & \dots & \rho^{|2-d|} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{|d-1|} & \rho^{|d-2|} & \dots & 1 \end{pmatrix}.$$

Dans ce cas, comme le montre la Figure 2.5, représentant le modèle graphique tel qu'estimé à l'aide de la technique *glasso* pour des variables auto-corrélées avec $\rho = 0.4$ et un paramètre de pénalisation arbitraire $\lambda = 0.1$, les seules variables reliées par une arête sont les variables X_i et X_j telles que $|i - j| = 1$. En effet, plus la distance entre i et j sera grande, moins la corrélation entre les variables X_i et X_j sera élevée. Ainsi, pour une valeur

adéquate de λ, seules les arêtes reliant deux variables adjacentes seront présentes dans le graphe.

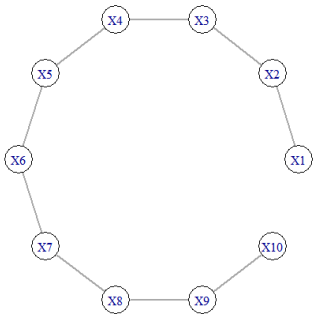

FIGURE 2.5 – Modèle graphique pour des variables auto-corrélées avec $\rho = 0.4$ et $\lambda = 0.1$.

Le graphique de la Figure 2.6, représentant la variation du nombre d'arêtes en fonction de la valeur de λ, montre bien que, une fois que la valeur de λ est suffisamment élevée pour limiter le nombre d'arêtes, le graphe possédera 9 arêtes ; jusqu'au moment où λ sera trop élevé et toutes les arêtes seront supprimées. Tout comme pour le cas indépendant, le nombre d'arêtes conservées avec les critères AIC et BIC est trop élevé par rapport au nombre attendu.

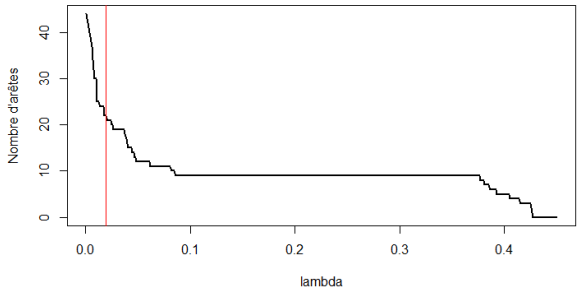

FIGURE 2.6 – Variation du nombre d'arêtes du graphe en fonction de λ, pour des variables auto-corrélées. La ligne rouge représente la valeur de λ qui minimise les critères AIC et BIC.

Cas par bloc

Dans ce cas, les variables situées dans le bloc sont équi-corrélées entre elles, alors que les autres ne le sont pas. La matrice Σ devient donc

$$\begin{pmatrix} \Sigma_{\rho,k} & 0_{k,d-k} \\ 0_{d-k,k} & I_{d-k} \end{pmatrix},$$

avec $1 \leq k \leq d$.

En considérant 10 variables dont seules les 3 premières sont corrélées ($k = 3$), la Figure 2.7, représentant le graphe obtenu à l'aide de la technique *glasso* avec $\rho = 0.4$ et $\lambda = 0.1$, montre que seules les arêtes reliant les variables X_1, X_2 et X_3 font partie du graphe. Ce nombre d'arêtes restera constant pour toutes les valeurs de λ pour autant que celui-ci ne soit pas trop faible (et ne supprime aucune arête) ou trop élevé (et ne les supprime toutes) ; ainsi qu'illustré à la Figure 2.8. Il est aussi possible de remarquer que si la valeur du paramètre de pénalisation était choisie en fonction des critères AIC et BIC alors, comme précédemment, le nombre d'arêtes conservées serait trop élevé au vu de la matrice de covariances réelle.

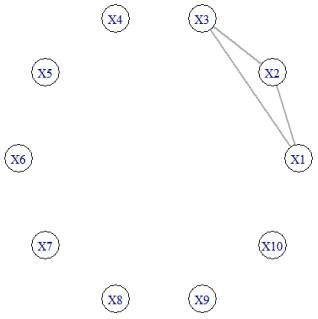

FIGURE 2.7 – Modèle graphique pour des variables corrélées par bloc, où seules les 3 premières variables sont corrélées, avec $\rho = 0.4$ et $\lambda = 0.1$.

Au vu de ces différents exemples, il est possible de conclure que la technique du *graphical lasso* estime correctement les paramètres du graphe lorsque la valeur du paramètre de pénalisation est bien choisie et que les données suivent une loi multinormale. Cependant, choisir la valeur de λ en fonction des critères AIC et BIC mène à un certain nombre d'arêtes conservées à tort et donc à une erreur d'estimation.

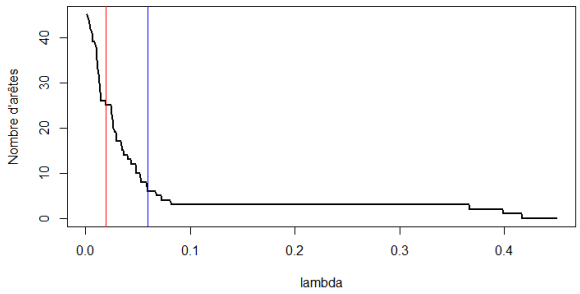

FIGURE 2.8 – Variation du nombre d'arêtes du graphe en fonction de λ, pour des variables avec une matrice de covariances par bloc. Les lignes rouge et bleue représentent les valeurs de λ qui minimisent les critères AIC et BIC respectivement.

Cas de données non standardisées

Les exemples numériques traités jusqu'à présent correspondaient à des données standardisées, ainsi que recommandé en théorie. Les deux exemples de cette section ne seront, quant à eux, pas standardisés. Ils vont ainsi permettre d'illustrer la différence entre les graphes créés sur des données standardisées ou non.

Premièrement, des données sont générées suivant une loi normale dont le vecteur moyen est le vecteur nul et dont la matrice de covariances est donnée par

$$\begin{pmatrix} \Sigma_{\rho,5} & 0_{5,5} \\ 0_{5,5} & 6\Sigma_{\rho,5} \end{pmatrix}. \tag{2.24}$$

Comme la corrélation entre les variables à l'intérieur des deux blocs est égale à ρ, il est attendu que le graphe forme deux cliques représentant les deux blocs et contienne donc 20 arêtes. Le graphique de la Figure 2.9 montre que le nombre d'arêtes des données non standardisées n'est pas si différent de ce qu'il devrait être après standardisation. Cependant un plus grand nombre d'arêtes sont conservées avant d'arriver à un graphe possédant 20 arêtes.

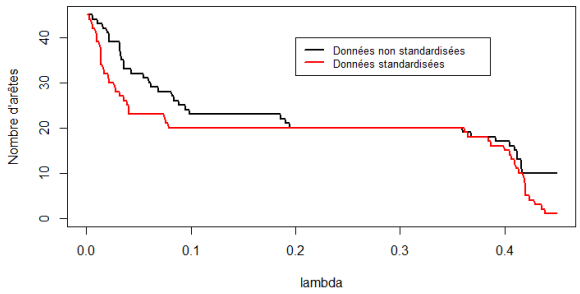

FIGURE 2.9 – Variation du nombre d'arêtes du graphe en fonction de λ, pour des variables avec la matrice de covariances (2.24) et pour la version standardisée des données (en rouge), avec $\rho = 0.4$.

Deuxièmement, des données sont générées suivant une loi normale dont le vecteur moyen est le vecteur nul et dont la matrice de covariances est donnée par

$$\begin{pmatrix} \Sigma_{\rho,5} & 0_{5,5} \\ 0_{5,5} & 6\Sigma_{\rho/10,5} \end{pmatrix}. \tag{2.25}$$

Cette fois, le deuxième bloc est caractérisé par une corrélation bien plus faible ($\frac{\rho}{10}$) que celle observée dans le premier bloc. C'est pourquoi il est attendu que le graphe forme une clique représentant le bloc des variables X_1 à X_5 et contienne donc 10 arêtes. Dans ce cas-ci (Figure 2.10), le nombre attendu d'arêtes n'est conservé que lorsque $\lambda = 0.4$. Le graphe possèdera donc principalement trop d'arêtes lorsque $\lambda < 0.4$ et ensuite trop peu lorsque $\lambda > 0.4$.

La Figure 2.11 représente les modèles graphiques créés à partir des données non standardisées à gauche et standardisées à droite, pour des valeurs du paramètre de pénalisation $\lambda = 0.15$ en haut et $\lambda = 0.4$ en bas, avec $\rho = 0.4$. Ces valeurs du paramètre de pénalisation ont été choisies de sorte à obtenir un graphe possédant 10 arêtes pour les deux ensembles de données. Pour $\lambda = 0.15$, le graphe construit sur les données standardisées (en haut à droite) possèdera 10 arêtes. Cette figure met en évidence le fait que, pour le modèle graphique basé sur les données non standardisées, lorsque le nombre d'arêtes du graphe est le nombre attendu, ce ne sont pas les bonnes arêtes qui sont conservées.

D'après ces deux exemples, il est possible de conclure que les arêtes entre les variables sont davantage conservées si ces dernières ont une variance plus élevée. Cela se traduit par une zone constante à 20 arêtes pour les données non standardisées, observée sur la Figure 2.10.

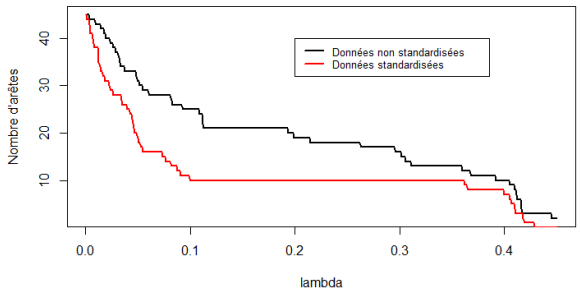

FIGURE 2.10 – Variation du nombre d'arêtes du graphe en fonction de λ, pour des variables avec la matrice de covariances (2.25) et pour la version standardisée des données (en rouge), avec $\rho = 0.4$.

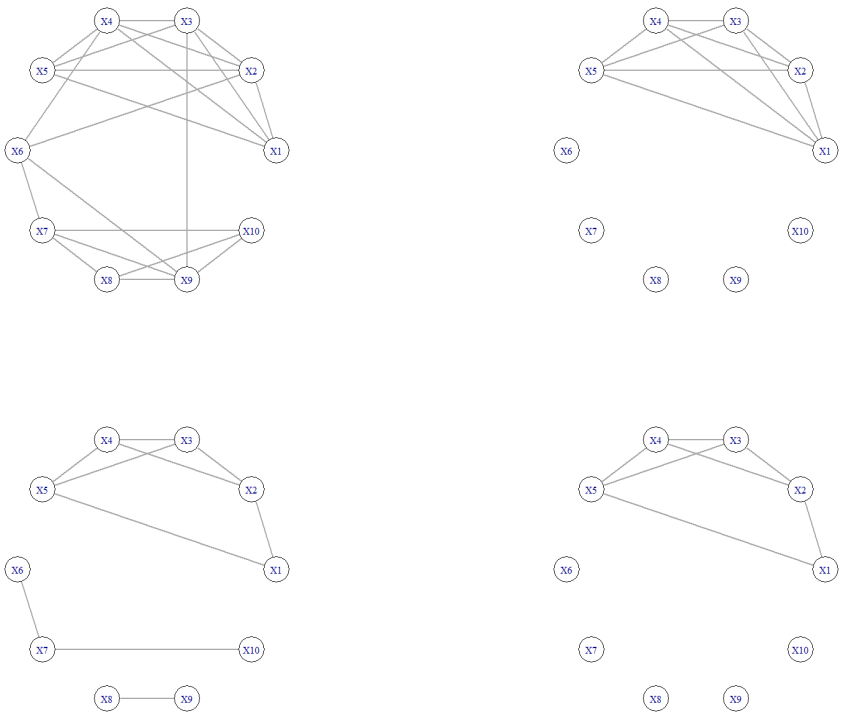

FIGURE 2.11 – Modèles graphiques représentant les données non standardisées à gauche et standardisées à droite, pour des valeurs du paramètre de pénalisation $\lambda = 0.15$ en haut et $\lambda = 0.4$ en bas, avec $\rho = 0.4$.

2.3.2 Exemples numériques lorsque la normalité n'est pas respectée

Dans cette section sont développés des exemples dont les données sont obtenues à partir de lois non normales. Ces exemples sont intéressants pour remarquer que l'hypothèse de normalité des données ne doit pas être violée.

Loi normale asymétrique

Pour ce premier exemple, c'est la loi normale asymétrique qui est étudiée. Cette loi est définie dans $\mathbb{R}^d$ par la fonction de densité suivante

$$p(x) = \phi_d(z, \Omega)\Phi(\alpha^T z)$$

où $\phi_d(z, \Omega)$ est la fonction de densité de la loi normale de dimension d, de moyenne nulle et de matrice de covariances Ω, Φ est la fonction de répartition de la loi $\mathcal{N}(0, 1)$ et α est un vecteur de dimension d. Le vecteur α est un paramètre de forme. Tout comme pour la loi multinormale, les dépendances conditionnelles entre les variables s'expriment à l'aide du paramètre Ω, mais aussi à l'aide du paramètre α. En effet,

$$X_i \amalg X_j | \{X_i, X_j\} \quad \Leftrightarrow \quad (\Omega^{-1})_{ij} = 0 \text{ et } \alpha_i \alpha_j = 0. \tag{2.26}$$

Plus d'informations sur la loi normale asymétrique sont détaillées dans [1], ainsi que sur les dépendances entre les variables dans [31].

Dans cet exemple, le paramètre Ω est défini en fonction de son inverse pour connaître les entrées de la matrice pour lesquelles Ω^{-1} est nul. La matrice Ω^{-1} a été choisie comme étant égale à

$$\Omega^{-1} = \begin{pmatrix} 1 & 0.4 & 0.4 & 0 & 0.4 & 0.4 & 0 & 0 & 0 & 0 \\ 0.4 & 1 & 0 & 0.4 & 0 & 0 & 0.4 & 0 & 0 & 0.4 \\ 0.4 & 0 & 1 & 0 & 0.4 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0.4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0.4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0 & 1 & 0.4 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0 & 0.4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.4 \\ 0 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 1 \end{pmatrix}. \tag{2.27}$$

Le paramètre α est, quant à lui, le vecteur $(2, 3, 1, 2, 0, 0, 0, 0, 0, 0)$.

Ainsi, au vu de ces paramètres et de (2.26), le graphe créé sur base des données générées à partir de cette loi normale multivariée asymétrique devrait être le graphe de gauche de la Figure 2.12. Or, après avoir généré 1000 données selon cette loi, le graphe estimé à l'aide

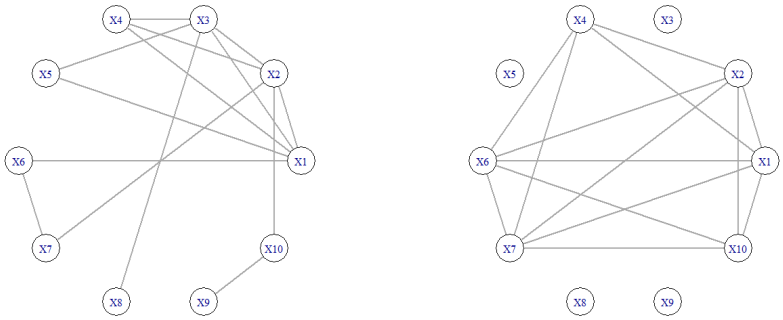

FIGURE 2.12 – A gauche : modèle graphique réel représentant les données générées à l'aide de la loi normale multivariée asymétrique. A droite : modèle graphique estimé à l'aide de la technique *glasso* sur ces mêmes données.

de la technique *glasso* possédant 14 arêtes, c'est à dire le nombre d'arêtes du graphe de gauche, n'est pas identique au graphe attendu. Le deuxième graphe est représenté à droite de la Figure 2.12.

Le graphe estimé possède donc certaines erreurs. Les arêtes entre les variables X_1 et X_3, X_1 et X_5, X_2 et X_3, X_3 et X_4, X_3 et X_5, X_3 et X_8 et X_9 et X_{10} ne sont pas conservées dans le graphe alors qu'elles le devraient et sont donc remplacées par d'autres arêtes. Ainsi, sur les 14 arêtes du graphe, 7 arêtes sont mal placées. Estimer la structure du graphe à l'aide de la technique *glasso* pour la loi normale asymétrique n'est donc pas convaincant.

Loi de puissance exponentielle

Après avoir obtenu les résultats pour une distribution asymétrique, il est intéressant de se demander ce qu'il se passe pour une distribution possédant un grand nombre de valeurs extrêmes, c'est à dire des queues épaisses, ou au contraire, des queues fines.

La loi de puissance exponentielle fait partie des distributions elliptiques. En effet, une fonction de répartition F de $\mathbb{R}^d$ est dite elliptique si sa fonction de densité f a la forme suivante :

$$f(x) = \det(S)^{-1/2} g[(x-\mu)^T S^{-1}(x-\mu)],$$

pour un vecteur $\mu \in \mathbb{R}^d$ et une matrice $d \times d$ symétrique et définie positive S. La fonction de densité de la loi de puissance exponentielle étant donnée par

$$f(x;\mu,\Sigma,\beta) = k|\Sigma|^{-1/2} \exp\left[\frac{-1}{2}[(x-\mu)^T\Sigma^{-1}(x-\mu)]^\beta\right],$$

avec

$$k = \frac{d\Gamma(d/2)}{\pi^{d/2}\Gamma(1+\frac{d}{2\beta})2^{1+\frac{d}{2\beta}}},$$

c'est bien une distribution elliptique avec $g(t) = \exp(\frac{-1}{2}t^{\beta})$. Cette distribution est détaillée dans [9].

Pour les distributions elliptiques, la matrice S^{-1} est appelée la pseudo matrice de concentration de F. Une simple transformation de cette matrice permet d'obtenir la matrice de corrélation partielle :

$$P = -BS^{-1}B$$

où B est la matrice diagonale $d \times d$ dont les éléments diagonaux sont égaux aux éléments diagonaux de S^{-1} à la puissance $\frac{-1}{2}$. Cette transformation est obtenue à l'aide de l'article [29]. La matrice P permet alors de construire le graphe des corrélations partielles. Ce graphe ne sera pas identique au graphe représentant les dépendances entre les variables. En effet, si $P_{ij} = 0$, cela signifie qu'il n'y a pas de dépendance linéaire entre les variables X_i et X_j, or une dépendance autre que linéaire pourrait exister.

Il est intéressant de regarder sur un exemple si le graphe obtenu par la méthode *glasso* représente bien la matrice de corrélation partielle P. Pour cet exemple, 1000 données sont générées selon une loi power-exponentielle de dimension 10, dont le vecteur μ est le vecteur nul, la matrice S est la matrice

$$S = \begin{pmatrix} \Sigma_{0.4,3} & 0_{3,2} & 0_{3,5} \\ 0_{2,3} & I_2 & 0_{2,5} \\ 0_{5,3} & 0_{5,2} & \Sigma_{0.6,5} \end{pmatrix}$$

et le paramètre β est choisi comme étant égal à 0.2 pour un premier exemple et à 20 pour un deuxième. Lorsque $\beta = 1$, la distribution de puissance exponentielle devient la distribution multinormale. Lorsque $\beta < 1$, les queues de la distribution seront plus fines que celles de la distribution normale, alors qu'elles seront plus épaisses lorsque $\beta > 1$.

La matrice P obtenue par transformation de S^{-1} est alors la matrice

$$\begin{pmatrix} -1 & 0.2857 & 0.2857 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2857 & -1 & 0.2857 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2857 & 0.2857 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0.2143 & 0.2143 & 0.2143 & 0.2143 \\ 0 & 0 & 0 & 0 & 0 & 0.2143 & -1 & 0.2143 & 0.2143 & 0.2143 \\ 0 & 0 & 0 & 0 & 0 & 0.2143 & 0.2143 & -1 & 0.2143 & 0.2143 \\ 0 & 0 & 0 & 0 & 0 & 0.2143 & 0.2143 & 0.2143 & -1 & 0.2143 \\ 0 & 0 & 0 & 0 & 0 & 0.2143 & 0.2143 & 0.2143 & 0.2143 & -1 \end{pmatrix}.$$

Au vu de cette matrice, le graphe représentant les données devrait être composé de deux cliques, une entre les 3 premières variables ainsi qu'une entre les 5 dernières. C'est, en effet,

le cas comme le montre le graphe de la Figure 2.13 représentant le graphe obtenu à l'aide de la technique *glasso*, pour $\lambda = 0.1$. Ce graphe est similaire, que $\beta = 0.2$, $\beta = 20$ ou qu'il soit créé sur base de la matrice P. En effet, au vu de la Figure 2.13, le nombre d'arêtes se stabilise à 13 pour les 3 exemples.

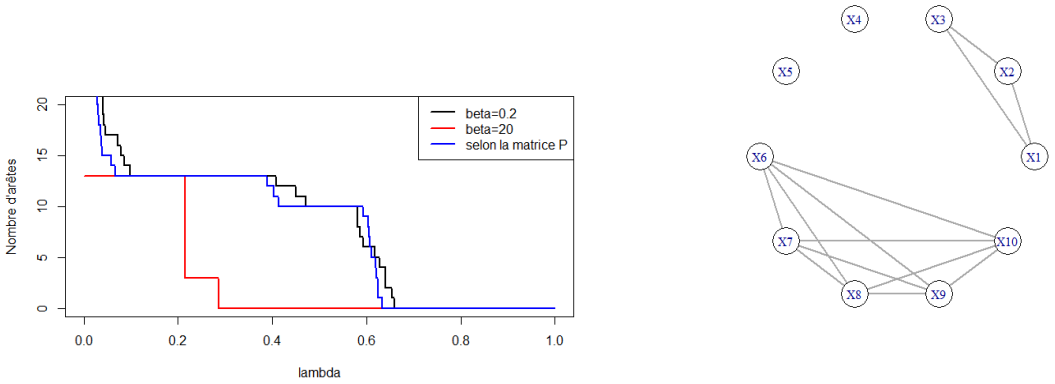

FIGURE 2.13 – A gauche : variation du nombre d'arêtes dans les modèles obtenus à l'aide de la technique *glasso* pour des données générées selon une loi de puissance exponentielle pour $\beta = 0.2$ et $\beta = 20$, ainsi que lorsque la technique est utilisée sur la matrice P. A droite : modèle graphique obtenu pour $\lambda = 0.1$, ce modèle étant identique pour les 3 exemples.

Lorsque la distribution des données fait partie de l'ensemble des distributions elliptiques, la matrice de corrélation des données est parfaitement représentée par la technique *glasso*. Cependant les dépendances non linéaires ne sont pas prises en compte.

Dépendances non linéaires

Comme vu dans l'exemple précédent, la technique *glasso* peut fonctionner correctement pour des fonctions de densité différentes de la loi normale, mais seules les dépendances linéaires sont prises en compte. Il est donc intéressant de se pencher sur le cas des dépendances non linéaires.

Pour cet exemple, 2 variables contenant 1000 données sont générées à partir des lois exponentielles et uniformes et 6 autres variables sont obtenues par transformation de ces deux premières :

— $X_1 \leftarrow \exp(1)$
— $X_2 \leftarrow X_1^2$
— $X_3 \leftarrow \sqrt{X_1}$
— $X_4 \leftarrow$ unif $(-2, 2)$
— $X_5 \leftarrow X_4^2$
— $X_6 \leftarrow X_4^3$
— $X_7 \leftarrow X_3 + X_5$
— $X_8 \leftarrow X_4 X_1$

Certaines dépendances entre ces variables sont proches d'une dépendance linéaire alors que d'autres pas du tout, comme la dépendance entre les variables X_4 et X_5 qui est clairement quadratique. En Figure 2.14 se trouve une représentation graphique de la matrice de corrélation pour ces variables.

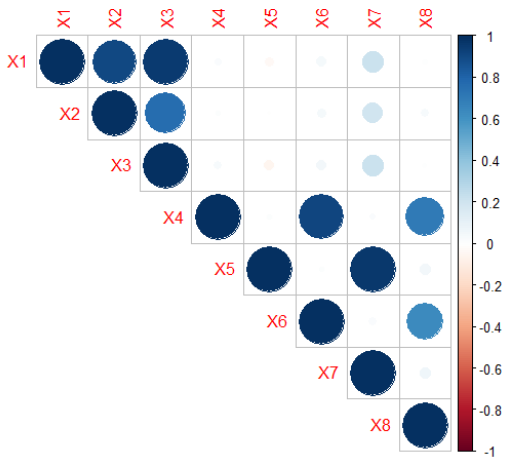

FIGURE 2.14 – Représentation de la matrice de corrélation pour les variables créées avec des dépendances non linéaires.

Au vu de la Figure 2.15, représentant la variation du nombre d'arêtes dans le graphe estimé par la méthode *glasso* pour les variables ci-dessus, représenter un graphe possédant 7 arêtes semble judicieux. En effet, le nombre d'arêtes dans le graphe reste constant à la valeur 7 pour plusieurs valeurs du paramètre de pénalisation.

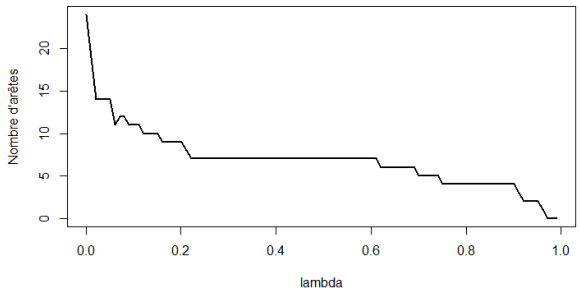

FIGURE 2.15 – Variation du nombre d'arêtes dans le graphe estimé à l'aide de la méthode *glasso* pour les variables créées avec des dépendances non linéaires.

Le graphe représentant ces variables, possédant 7 arêtes et obtenu à l'aide de la technique *glasso* est représenté en Figure 2.16. Il est clair que ce graphe ne représente que les dépendances linéaires entre les variables. En effet, les arêtes comprises dans le graphe ne lient que les couples de variables dont la corrélation est forte. Les couples de variables comme (X_4, X_5) et (X_5, X_6) ayant une forte dépendance non linéaire ne sont, quant à eux, pas reliés dans le graphe.

Il semble donc, au vu de ces deux derniers exemples, que seules les dépendances linéaires sont exprimées dans les modèles graphiques obtenus par la méthode *glasso*, lorsque la fonction de densité des données n'est pas la densité multinormale. Cette conclusion parait logique car la méthode *glasso* se base uniquement sur la matrice de covariances des données.

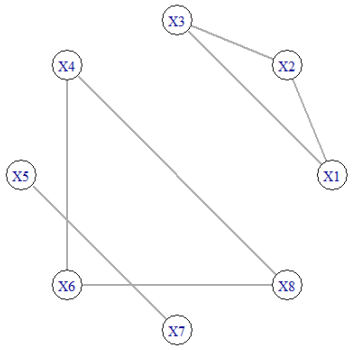

FIGURE 2.16 – Modèle graphique tel qu'estimé à l'aide de la méthode *glasso* pour les variables créées avec des dépendances non linéaires, avec $\lambda = 0.4$.

2.3.3 Modèles graphiques *glasso* des données socio-économiques de la Wallonie

La présentation théorique de la technique *glasso* repose sur la normalité. Afin de construire le graphe relatif aux données socio-économiques, il faudrait donc, avant tout, vérifier si la normalité est acceptable. Le test de Mardia permet de vérifier si la normalité des données est acceptable en dimension d. Pour les variables expliquant l'économie des communes wallonnes, le test de Mardia rejette la normalité. Le graphe produit par la théorie des modèles graphiques gaussiens pourrait donc posséder un certain nombre d'arêtes mal placées. Ce graphe sera tout de même représenté pour pouvoir le comparer avec les graphes créés à partir d'autres méthodes ne se basant pas sur la normalité.

En Figure 2.17 se trouve une visualisation graphique de la matrice de corrélation de la base de données. Plus le point représenté est gros, plus les variables sont corrélées. Des variables corrélées positivement, c'est-à-dire avec une corrélation entre 0 et 1 seront représentées par un point bleu tandis que les variables corrélées négativement seront représentées par un point rouge. Ainsi, quand la valeur de λ diminue, seules les arêtes entre les variables dont la corrélation en valeur absolue est la plus grande seront conservées.

Comme expliqué précédemment, la minimisation des critères AIC et BIC peut être utilisée pour trouver la meilleure valeur du paramètre λ de la pénalisation lasso. Cependant, ces deux critères sont strictement croissants comme le montre la Figure 2.18 représentant la variation du critère BIC en fonction du paramètre de pénalisation λ.

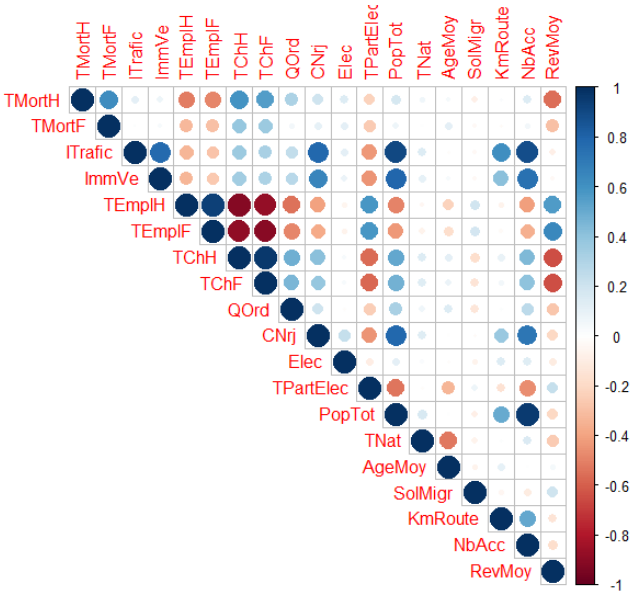

FIGURE 2.17 – Représentation de la matrice de corrélation de la base de données.

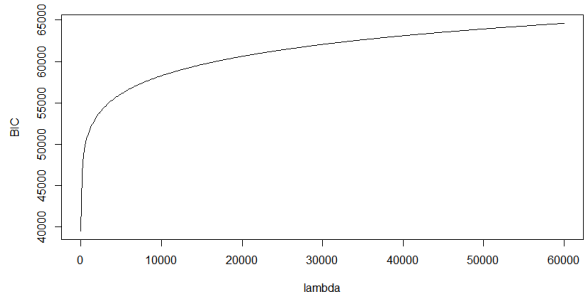

FIGURE 2.18 – Variation du critère BIC en fonction du paramètre de pénalisation λ.

Le fait que la valeur du paramètre de pénalisation minimisant le critère BIC soit égal à 0 est un peu interpellant. Pour se conforter dans l'idée qu'obtenir une valeur nulle du paramètre n'arrive que lorsque la matrice de covariances des données ne possède pas de valeur nulle, des simulations ont été effectuées. D'après celles-ci, lorsque la matrice de concentration possède plusieurs valeurs nulles, le paramètre ne sera jamais nul. Quand la matrice de concentration des données ne possède aucune valeur nulle, le paramètre de pénalisation tel que suggéré par ces critères sera, la plupart du temps, nul. Ces deux conclusions sont bien celles attendues. Ainsi, il semblerait, d'après les critères AIC et BIC, que toutes les variables de la base de données soient dépendantes. Cependant, il est intéressant de ne conserver dans le graphe, que les arêtes entre les variables dont la dépendance est la plus forte.

Comme ces deux critères ne sont pas exploitables dans le cas de la base de données, un choix particulier du nombre d'arêtes doit être fait de sorte à trouver les valeurs de λ correspondantes. Après avoir effectué quelques recherches, aucune indication concernant un nombre adéquat d'arêtes n'est mentionné dans la littérature. Pour limiter le nombre d'arêtes dans le graphe, un choix personnel doit donc être fait ou alors une structure particulière peut être imposée au graphe (voir chapitre 4). Un premier choix, en relation avec le chapitre 4 où il sera question de construire une forêt, est de créer un graphe possédant 18 arêtes. Cela vient du fait qu'une forêt possède au maximum $d - 1 = 18$ arêtes. Pour faire un autre choix de paramètre, il faut principalement se baser sur la connaissance à priori des dépendances entre les variables. Dans le chapitre 1, certaines relations de dépendance à priori ont déjà été relevées, comme par exemple, la dépendance entre les taux de chômage et d'emploi pour les hommes et les femmes. De plus, la matrice de corrélation reprend l'ensemble des dépendances linéaires entre toutes les variables de la base de données. C'est pourquoi le deuxième choix effectué est de limiter le graphe à 36 arêtes ; c'est le nombre de couples de variables ayant une corrélation supérieure, en valeur absolue, à 0.5. Ainsi, seules les couples de variables les plus corrélées, c'est-à-dire les couples ayant les points les plus gros dans la Figure 2.17, devraient être reliés par une arête.

Les deux graphes, possédant 36 et 18 arêtes, représentant la base de données sont construits en utilisant la technique du *graphical lasso* pour $\lambda = 41300$ et $\lambda = 4100$ respectivement. La variation du nombre d'arêtes du graphe en fonction de λ est représentée en Figure 2.19. Lorsque λ augmente, le nombre d'arêtes doit diminuer. Il est cependant possible que la courbe ne soit pas toujours décroissante car, pour chaque valeur de λ, l'algorithme *glasso* est recalculé et peut prendre des chemins différents pour arriver à la convergence. Une petite variation dans le nombre d'arêtes peut donc être observée.

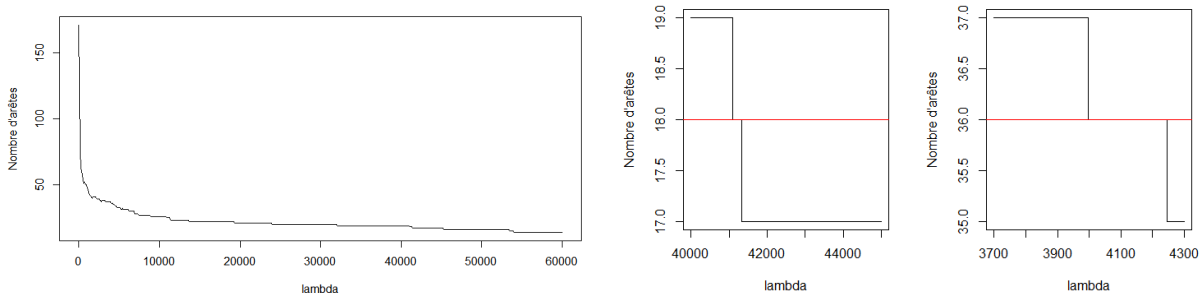

FIGURE 2.19 – Variation du nombre d'arêtes du graphe en fonction du paramètre de pénalisation λ.

Avant de créer un graphe à partir de la base de données, il faut rappeler que la technique *graphical lasso* n'est pas équivariante. Il est donc nécessaire de standardiser les données avant de commencer.

Les graphes représentant la base de données non standardisée sont tout de même représentés pour les comparer aux autres graphes. Ceux-ci sont repris en Figure 2.20.

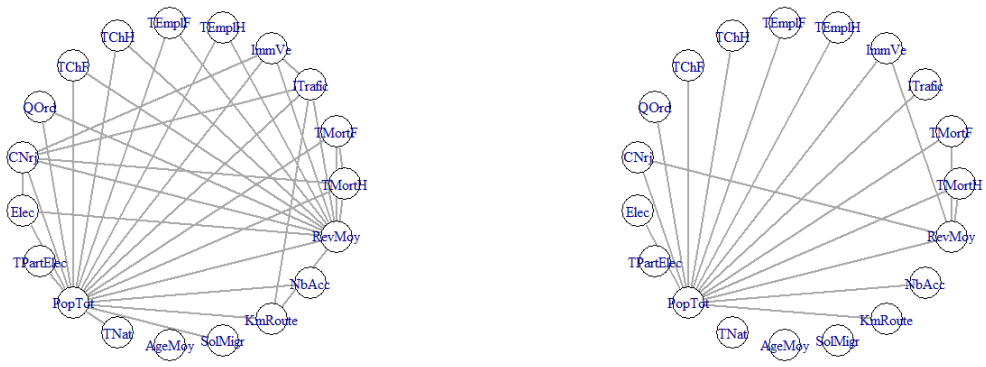

FIGURE 2.20 – Modèles graphiques pour la base de données contenant 36 arêtes à gauche et 18 arêtes à droite.

Sur base du graphe de la Figure 2.20, il est possible de voir que les graphes ne respectent pas tellement les corrélations entre les variables telles qu'illustrées à la Figure 2.17. En effet, la plupart des arêtes sont liées aux variables PopTot et RevMoy. Cela est du au fait que les variances des variables PopTot et RevMoy sont largement plus élevées que la variance des autres variables. Si les variables sont centrées et réduites, les modèles graphiques deviennent les modèles repris en Figure 2.21. Ceux-ci respectent davantage les corrélations entre les variables. Ainsi il est possible de voir que le graphe de droite possède une clique contenant les variables TChF, TChH, TEmplF et TEmplH. Cela confirme qu'il semble judicieux de penser que les taux de chômage et d'emploi sont liés. Il est possible de constater aussi que les variables concernant le trafic routier sont fortement dépendantes du nombre d'habitants de la commune et de la consommation d'énergie.

Le graphe de gauche de la Figure 2.21 représente assez bien la matrice de corrélation. Seules deux arêtes entre deux couples de variables sont conservées alors que leur corrélation est, en valeur absolue, inférieure à 0.5. C'est le cas des couples QOrd et TchH ainsi que PopTot et TEmplH. Ces deux arêtes sont conservées à la place des arêtes entre les variables RevMoy et TEmplH et entre KmRoute et PopTot.

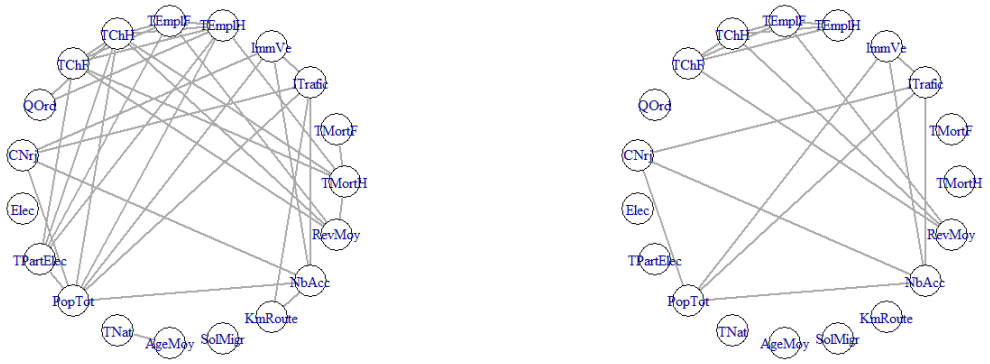

FIGURE 2.21 – Modèles graphiques pour la base de données dont les variables ont été centrées et réduites, contenant 36 arêtes à gauche et 18 arêtes à droite.

Comme remarqué précédemment, lorsque la technique *glasso* est appliquée à des données dont la distribution jointe n'est pas gaussienne, le graphe obtenu ne semble uniquement tenir compte des dépendances linéaires entre les variables. Dans le cas des données socio-économiques de la Wallonie, les dépendances linéaires sont bien reprises dans le graphe, même si quelque erreurs sont commises. Cependant, il est impossible de savoir, avec ces graphes, si d'autres dépendances non linéaires apparaissent entre les variables. Pour le découvrir, il est nécessaire d'appliquer à ces données une technique d'estimation qui ne demande pas d'hypothèse de normalité. Une technique de la sorte est développée dans le chapitre suivant.

Chapitre 3

Modèles graphiques du type non paranormal

Pour élargir les hypothèses des modèles graphiques, il est possible de considérer un vecteur non gaussien. Dans ce chapitre, des hypothèses non paramétriques vont être supposées. Ce chapitre est rédigé en se basant sur les articles [18] et [21].

La forme générale pour une densité de probabilité (strictement positive) décrite par un graphe non dirigé G est

$$p(x) = \frac{1}{k} \prod_{C \in \text{Cliques}(G)} \psi_C(x_C), \tag{3.1}$$

qui peut se réécrire sous la forme

$$p(x) = \frac{1}{k} \exp\left(\sum_{C \in \text{Cliques}(G)} f_C(x_C) \right), \tag{3.2}$$

où la somme est prise sur toutes les cliques du graphe. En général, ce type de graphe est nommé *un modèle graphique non paramétrique*. C'est le modèle graphique analogue au modèle général de régression non paramétrique. Le modèle est constitué de deux éléments principaux, le graphe et les fonctions $\{f_C\}$. La difficulté majeure pour travailler avec un tel modèle est la constante de normalisation k qui, en général, ne peut pas être calculée ou approchée efficacement. Par ailleurs, afin de travailler avec le modèle (3.2), deux types d'hypothèses peuvent être considérées ; soit imposer une structure aux fonctions f_C, soit l'imposer au graphe.

Seule la première approche est détaillée dans ce chapitre. Celle-ci consiste à remplacer le vecteur aléatoire $X = (X_1, \ldots, X_d)$ par le vecteur aléatoire transformé $f(X) = (f_1(X_1), \ldots, f_d(X_d))$, de manière à retomber sur une loi gaussienne multivariée. Cela se traduit par une extension non paramétrique de la loi normale, appelée distribution *non paranormale*. Le non paranormal dépend des fonctions univariées f_j, d'une moyenne μ et d'une matrice de covariance Σ, lesquelles sont habituellement estimées à partir des données. Alors que la famille de distribution résultante est plus riche que celle de la normale

paramétrique standard, la relation d'indépendance entre les variables est toujours reprise dans la matrice de concentration $\Theta = \Sigma^{-1}$. Bien entendu, pour pouvoir exploiter le vecteur aléatoire transformé, il faut déterminer les fonctions f_j.

Avant de décrire l'approche non paranormale pour la définition d'un modèle graphique, la distribution non paranormale va être définie en toute généralité.

3.1 Définition

Un vecteur aléatoire $X = (X_1, \ldots, X_d)$ a une distribution *non paranormale*, notée

$$X \sim NPN(\mu, \Sigma, f), \tag{3.3}$$

où Σ est d-carrée et définie positive, s'il existe des fonctions $\{f_j\}_{j=1}^d$ telles que

$$Z \equiv f(X) \sim \mathcal{N}_d(\mu, \Sigma), \text{ où } f(X) = (f_1(X_1), \ldots, f_d(X_d)).$$

Quand les f_j sont monotones et différentiables, la fonction de densité de probabilité jointe de X est donnée par

$$p(x_1, \cdots, x_d) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(f(x) - \mu)^T \Sigma^{-1}(f(x) - \mu)\} \prod_{j=1}^{d} |f_j^{'}(x_j)|, \tag{3.4}$$

où le terme produit est le jacobien de la transformation.

La densité en (3.4) n'est pas identifiable. En effet, toutes les fonctions peuvent être multipliées par une constante, ainsi que μ et la diagonale de Σ, sans changer la densité. Pour que la famille soit identifiable, il est demandé que les f_j préservent les moyennes et variances marginales :

$$\mu_j = \mathbb{E}[Z_j] = \mathbb{E}[X_j] \text{ et } \sigma_j^2 \equiv \Sigma_{jj} = \mathrm{Var}(Z_j) = \mathrm{Var}(X_j). \tag{3.5}$$

Ces conditions ne dépendent que de la diagonale de Σ et pas de toute la matrice de covariances.

Illustration personnelle

En Figure 3.1 se trouve un exemple de densité d'une loi non paranormale en dimension 2. Les fonctions utilisées sont $f_1(x_1) = \mathrm{sign}(x_1)\sqrt{|x_1|}$ et $f_2(x_2) = x_2^3$. Le vecteur moyen est le vecteur nul et la matrice de covariances est donnée par

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

avec $-1 \leq \rho \leq 1$. Le vecteur aléatoire $Z = (Z_1, Z_2) = \Big(\text{sign}(X_1)\sqrt{|X_1|}, X_2^3\Big)$ suit une loi binormale de moyenne nulle et de matrice de covariances Σ. La fonction de densité du couple (X_1, X_2) reprise en (3.4) devient alors

$$p(x_1, x_2) = \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2}\frac{1}{1-\rho^2}\Big((\text{sign}(x_1)\sqrt{|x_1|})^2 - 2\rho x_2^3 \text{sign}(x_1)\sqrt{|x_1|} + x_2^6\Big)\right\} \left(\frac{3}{4}\frac{1}{\sqrt{|x_1|}}x_2^2\right).$$

La Figure 3.1 représente cette densité pour des valeurs de ρ égales à 0, 0.5 et -0.8. Lorsque la valeur du paramètre de corrélation n'est pas nulle, les deux pics centrés en $(0, 1)$ et $(0, -1)$ ne sont plus symétriques.

3.2 Choix des fonctions de transformation

Soit $X = (X_1, \cdots, X_d)$ un vecteur aléatoire. Supposons que le vecteur X soit de loi non paranormale et que la distribution marginale de X_j soit $F_j(x)$. Alors, ainsi que mis en évidence dans la propriété suivante, il existe un lien explicite entre f_j et F_j.

Proposition 2. *Si $X \sim NPN(\mu, \Sigma, f)$ et si $X_j \sim F_j$ pour $j = 1, \cdots, d$, alors*

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)), \tag{3.6}$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Démonstration. Comme le composant $f_j(X_j) = Z_j$ est gaussien de moyenne μ_j et de variance σ_j^2, il advient que

$$F_j(x) = \mathbb{P}(X_j \leq x) = \mathbb{P}(Z_j \leq f_j(x)) = \Phi\Big(\frac{f_j(x) - \mu_j}{\sigma_j}\Big). \tag{3.7}$$

Ce qui implique que

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)). \tag{3.8}$$

□

Il faudrait maintenant montrer que, si $X \sim F$, avec $X_j \sim F_j$ et si on définit f_j comme dans la proposition 2, alors $X \sim NPN(\mu, \Sigma, f)$. Pour cela, il est nécessaire d'introduire la notion de copule [24].

Tout d'abord, rappelons que si F_j est la fonction de répartition de X_j, alors $U_j = F_j(X_j)$ est distribué uniformément sur $[0, 1]$. Si C dénote la fonction de répartition jointe

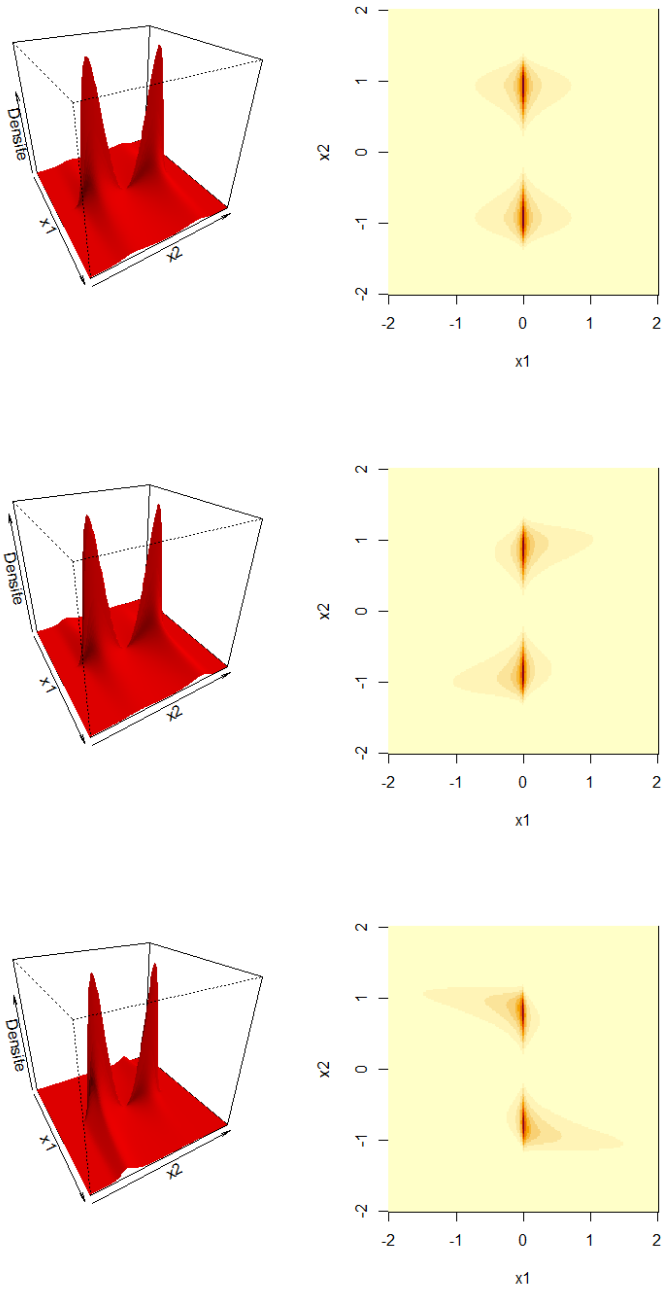

FIGURE 3.1 – Densité d'une loi non paranormale dont les fonctions de transformation sont données par $f_1(x) = \text{sign}(x)\sqrt{|x|}$ et $f_2(x) = x^3$, pour différentes valeurs de la corrélation ($\rho = 0$ pour la première ligne, $\rho = 0.5$ pour la deuxième et $\rho = -0.8$ pour la troisième).

de $U = (U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$, et si F dénote la fonction de répartition de X, alors

$$\begin{aligned} F(x_1, \ldots, x_d) &= \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ &= \mathbb{P}(F_1(X_1) \leq F_1(x_1), \ldots, F_d(X_d) \leq F_d(x_d)) \\ &= \mathbb{P}(U_1 \leq F_1(x_1), \ldots, U_d \leq F_d(x_d)) \\ &= C(F_1(x_1), \ldots, F_d(x_d)). \end{aligned}$$

Ce développement est connu sous le nom de *théorème de Sklar* [24], et la fonction C est une copule. Si les fonctions de répartition marginales sont continues, alors la copule C est unique. Si C admet la densité c, alors

$$p(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{j=1}^{d} p_j(x_j), \tag{3.9}$$

où $p_j(x_j)$ est la densité marginale de X_j et où $p(x_1, \ldots, x_d)$ est la fonction de densité de X.

Il est maintenant possible de considérer la proposition suivante.

Proposition 3. *Soit $X \sim F$, avec $X_j \sim F_j$ pour $j = 1, \cdots, d$. Si $f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x))$, alors $X \sim NPN(\mu, \Sigma, f)$ car $Z = f(X) \sim \mathcal{N}_d(\mu, \Sigma)$.*

Démonstration. Pour vérifier que les fonctions f_j conviennent, c'est-à-dire pour retomber sur la définition d'une loi non paranormale, il faut montrer que $Z = f(X)$ est distribué selon une loi normale de dimension d.

Grâce à (3.9), il vient

$$p(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{j=1}^{d} p_j(x_j).$$

En effectuant le changement de variables $z_j = f_j(x_j) = \mu_j + \sigma_j \Phi^{-1}(F_j(x_j))$, dont le jacobien est une matrice diagonale dont les éléments diagonaux sont donnés par

$$(\mathcal{J})_{jj} = \frac{\phi(\frac{y_j - \mu_j}{\sigma_j}) \frac{1}{\sigma_j}}{p_j(F_j^{-1}(\Phi(\frac{y_j - \mu_j}{\sigma_j})))}, \tag{3.10}$$

où ϕ fait respectivement référence à la fonction de densité de la loi normale centrée réduite et où l'inverse des fonctions F_j est défini par $F_j^{-1}(t) = \inf\{x : F(x) \geq t\}$. La fonction de

densité de $(Z_1, \cdots, Z_d)$, notée $g(z_1, \cdots, z_d)$ devient

$$\begin{aligned} g(z_1, \cdots, z_d) &= c\Big(\Phi(\frac{z_1-\mu_1}{\sigma_1}), \cdots, \Phi(\frac{z_d-\mu_d}{\sigma_d})\Big) \prod_{j=1}^{d} p_j\Big(F_j^{-1}(\Phi(\frac{z_j-\mu_j}{\sigma_j}))\Big) \prod_{i=1}^{d} (\mathcal{J})_{ii} \\ &= c\Big(\Phi(\frac{z_1-\mu_1}{\sigma_1}), \cdots, \Phi(\frac{z_d-\mu_d}{\sigma_d})\Big) \prod_{j=1}^{d} \phi\Big(\frac{z_j-\mu_j}{\sigma_j}\Big)\frac{1}{\sigma_j} \\ &= c\Big(\Phi_{\mu_1,\sigma_1}(z_1), \cdots, \Phi_{\mu_d,\sigma_d}(z_d)\Big) \prod_{j=1}^{d} \phi_{\mu_j,\sigma_j}(z_j). \end{aligned}$$

Or, comme la copule est invariante si les variables sont modifiées par une fonction strictement croissante (voir page 25 du livre [24]), cette expression correspond exactement à la densité jointe d'une loi normale multivariée de moyenne μ et de matrice de covariances Σ. □

Illustration personnelle

Voici maintenant des illustrations de ce changement de variables en deux dimensions.

Le premier exemple consiste à considérer des variables indépendantes. Si les variables X_i, $i = 1, 2$, sont de moyennes et variances respectives μ_i et σ_i^2, et si leurs fonctions de densité et de répartition sont notées p_i et F_i respectivement, alors la fonction de densité jointe du vecteur $X = (X_1, X_2)$ est, par indépendance,

$$p(x_1, x_2) = p_1(x_1)p_2(x_2). \tag{3.11}$$

Si la variable X_1 est distribuée selon une loi exponentielle de moyenne 1 et si la variable X_2 est distribuée selon la loi uniforme sur $[1, 5]$, alors

$$p_1(x) = e^{-x}\chi_{[0,+\infty[}(x), \quad F_1(x) = (1-e^{-x})\chi_{[0,+\infty[}(x), \quad \mathbb{E}[X_1] = 1 = \mathrm{Var}[X_1],$$

$$p_2(x) = \frac{1}{4}\chi_{[1,5]}(x), \quad F_2(x) = \frac{x-1}{4}\chi_{[1,5]}(x), \quad \mathbb{E}[X_2] = 3, \quad \mathrm{Var}[X_2] = \frac{4}{3}.$$

Ainsi, si $X_1 \amalg X_2$, la densité jointe de (X_1, X_2) est

$$p(x_1, x_2) = \frac{\exp(-x_1)}{4}\chi_{[1,5]}(x_2)\chi_{[0,+\infty[}(x_1). \tag{3.12}$$

La densité du vecteur aléatoire (X_1, X_2) est reprise en Figure 3.2.

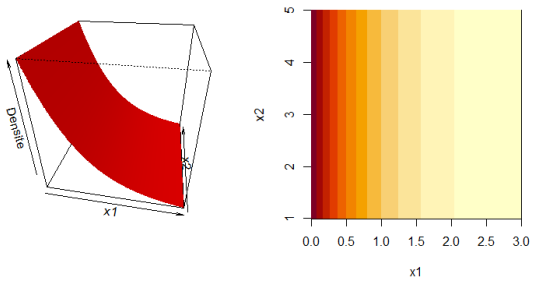

FIGURE 3.2 – Graphique et courbes de niveau de la fonction de densité du vecteur aléatoire (X_1, X_2) où $X_1 \sim \text{Exp}(1)$, $X_2 \sim \text{Unif}[1,5]$ et $X_1 \amalg X_2$.

Si maintenant, le changement de variables

$$\begin{cases} z_1 = \mu_1 + \sigma_1 \Phi^{-1}(F_1(x1)) = 1 + \Phi^{-1}(1 - e^{-x_1}) \\ z_2 = \mu_2 + \sigma_2 \Phi^{-1}(F_2(x2)) = 3 + \sqrt{\frac{4}{3}} \Phi^{-1}(\frac{x_2 - 1}{4}) \end{cases}$$

est effectué, il s'ensuit que

$$\begin{cases} x_1 = -\ln(1 - \Phi(z_1 - 1)) \\ x_2 = 1 + 4\Phi\Big(\sqrt{\frac{3}{4}}(z_2 - 3)\Big) \end{cases}$$

et le déterminant du jacobien est

$$\frac{\phi(z_1 - 1)}{1 - \Phi(z_1 - 1)} 4\sqrt{\frac{3}{4}} \phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right).$$

La densité du vecteur aléatoire $Z = (Z_1, Z_2) = (f_1(X_1), f_2(X_2))$ est donc donnée par

$$\begin{aligned} g(z_1, z_2) &= \frac{1 - \Phi(z_1 - 1)}{4} \chi_{[1,5]}\left(1 + 4\Phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right)\right) \frac{\phi(z_1 - 1)}{1 - \Phi(z_1 - 1)} 4\sqrt{\frac{3}{4}} \phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \\ &= \phi(z_1 - 1)\sqrt{\frac{3}{4}} \phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \chi_{[1,5]}\left(1 + 4\Phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right)\right) \\ &= \phi_{1,1}(z_1) \phi_{3,\frac{4}{3}}(z_2) \end{aligned}$$

car

$$\begin{aligned} & 1 \leq 1 + 4\Phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \leq 5 \\ \Leftrightarrow & 0 \leq \Phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \leq 1 \end{aligned}$$

ce qui est toujours vrai. Cette densité est l'expression d'une fonction de densité normale centrée en $(1, 3)$ et dont la matrice de covariances est la matrice diagonale

$$\begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 4 \end{pmatrix}. \tag{3.13}$$

Cela permet de conclure que le vecteur X suit une loi non paranormale, ainsi sa fonction de densité peut aussi être obtenue à l'aide de la transformation (3.4).

Si maintenant les variables sont corrélées, alors la matrice de covariances du vecteur $X = (X_1, X_2)$ est donnée par

$$\begin{pmatrix} 1 & \rho \\ \rho & \frac{4}{3} \end{pmatrix}, \tag{3.14}$$

où $\rho < \sigma_1\sigma_2$. En utilisant le résultat de la densité non paranormale, la densité jointe de (X_1, X_2) devient alors

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{1}{1-\rho^2}\begin{pmatrix} f_1(x_1)-\mu_1 \\ f_2(x_2)-\mu_2 \end{pmatrix}^T \begin{pmatrix} \frac{1}{4} & \rho \\ \rho & 5 \end{pmatrix}\begin{pmatrix} f_1(x_1)-\mu_1 \\ f_2(x_2)-\mu_2 \end{pmatrix}\right\} \prod_{i=1}^{2} \frac{p_i(x_i)\sigma_i}{\phi(\Phi^{-1}(F_i(x_i)))} \\ &= \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}} \frac{p_1(x_1)}{\phi(\Phi^{-1}(F_1(x_1)))}\frac{p_2(x_2)}{\phi(\Phi^{-1}(F_2(x_2)))} \\ &\quad \exp\left\{-\frac{1}{2}\frac{1}{1-\rho^2}\Big((\Phi^{-1}(F_1(x_1)))^2 - 2\rho\Phi^{-1}(F_1(x_1))\Phi^{-1}(F_2(x_2)) + (\Phi^{-1}(F_2(x_2)))^2\Big)\right\}. \end{aligned}$$

Cette densité est représentée pour $\rho = 0.1$ et $\rho = -0.1$ en Figure 3.3. Pour des ρ dont la valeur absolue est plus grande que 0.1, le pic augmente et la pente pour descendre de ce pic devient de plus en plus raide.

Pour montrer que le vecteur aléatoire $Z = f(X)$ est distribué normalement, la copule normale sera utilisée. Celle-ci est définie en fonction de la matrice de corrélation $R = \text{diag}(\sigma)^{-1}\Sigma\text{diag}(\sigma)^{-1}$ des variables. La densité de la copule normale est donnée par

$$\begin{aligned} c(u_1, u_2) &= \frac{1}{\sqrt{\det R}} \exp\left\{\frac{-1}{2}\begin{pmatrix} \Phi^{-1}(u_1) \\ \Phi^{-1}(u_2) \end{pmatrix}^T R^{-1} \begin{pmatrix} \Phi^{-1}(u_1) \\ \Phi^{-1}(u_2) \end{pmatrix}\right\} \frac{1}{\phi(\Phi^{-1}(u_1))}\frac{1}{\phi(\Phi^{-1}(u_2))} \\ &= \phi_R(\Phi^{-1}(u_1), \Phi^{-1}(u_2))\frac{1}{\phi(\Phi^{-1}(u_1))}\frac{1}{\phi(\Phi^{-1}(u_2))}. \end{aligned}$$

Comme la densité du vecteur aléatoire x est

$$\begin{aligned} p(x_1, x_2) &= c(F_1(x_1), F_2(x_2))p_1(x_1)p_2(x_2) \\ &= \phi_R(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)))\frac{p_1(x_1)}{\phi(\Phi^{-1}(F_1(x_1)))}\frac{p_2(x_2)}{\phi(\Phi^{-1}(F_2(x_2)))}, \end{aligned}$$

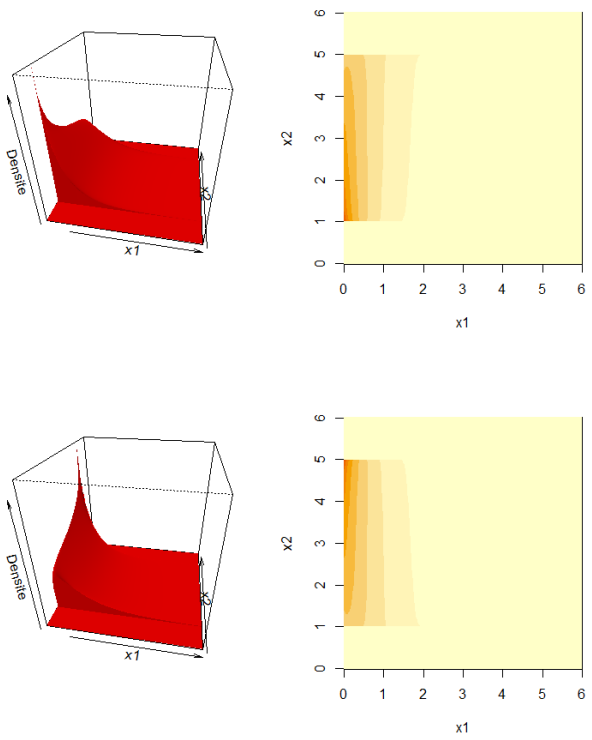

FIGURE 3.3 – Graphiques et courbes de niveau de la fonction de densité jointe du vecteur aléatoire (X_1, X_2) pour des valeurs de corrélation $\rho = 0.1$ (en haut) et $\rho = -0.1$ (en bas) avec $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}\,[1, 5]$.

la densité du vecteur aléatoire Z devient, en effectuant le même changement de variables que précédemment,

$$\begin{aligned} g(z_1, z_2) &= \phi_R\left(z_1 - 1, \sqrt{\frac{3}{4}}(z_2 - 3)\right) \frac{1 - \Phi(z_1 - 1)}{\phi(z_1 - 1)} \frac{1}{4} \frac{1}{\phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right)} \\ &\quad \chi_{[1,5]}\left(1 + 4\Phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \frac{\phi(z_1 - 1)}{1 - \Phi(z_1 - 1)} 4\phi\left(\sqrt{\frac{3}{4}}(z_2 - 3)\right) \sqrt{\frac{3}{4}}\right. \\ &= \phi_R\left(z_1 - 1, \sqrt{\frac{3}{4}}(z_2 - 3)\right) \sqrt{\frac{3}{4}} \\ &= \phi_{\mu,\Sigma}(z_1, z_2). \end{aligned}$$

Le vecteur Z suit bien une loi binormale dont le vecteur moyen et la matrice de covariances sont les mêmes que ceux du vecteur X.

3.3 Estimation

En pratique, les fonctions de répartition marginales des variables sont inconnues. Ainsi, pour pouvoir obtenir les fonctions f_j par l'expression (3.6), il est tout d'abord nécessaire d'estimer les fonctions de répartition. Pour ce faire, les deux choix assez naturels sont : soit supposer que les variables suivent une loi paramétrique dont les paramètres sont inconnus et estimer ces paramètres, soit estimer la fonction de répartition de manière non paramétrique. C'est cette dernière option qui est suggérée dans l'article [18] et qui est détaillée ci-dessous.

Soit $(X_{ij})_{1\leq i\leq n, 1\leq j\leq d}$ un échantillon de taille n où $X_i = (X_{i1}, \ldots, X_{id})^T \in \mathbb{R}^d$. Voici la procédure d'estimation non paramétrique des fonctions de répartition marginales.

Un candidat naturel pour estimer F_j est $\hat{F}_j$, la fonction de répartition marginale empirique

$$\hat{F}_j(t) = \frac{1}{n}\sum_{i=1}^{n} \chi_{\{X_{ij}\leq t\}}.$$

Cependant, la plus petite valeur de $\hat{F}_j$ est 0 et sa valeur la plus grande est 1, et dans ces deux cas, $\Phi^{-1}(\hat{F}_j)$ est infini. Dans le cas d'ensembles de données de grandes dimensions, c'est-à-dire quand d peut augmenter avec n, une alternative consiste à utiliser un estimateur tronqué ou estimateur *Winsorized*,

$$\tilde{F}_j(x) = \begin{cases} \delta_n & \text{si } \hat{F}_j(x) < \delta_n, \\ \hat{F}_j(x) & \text{si } \delta_n \leq \hat{F}_j(x) \leq 1-\delta_n, \\ 1-\delta_n & \text{si } \hat{F}_j(x) > 1-\delta_n, \end{cases} \tag{3.15}$$

où δ_n est un paramètre de troncature à fixer. Souvent, celui-ci est choisi tel que

$$\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi \ln n}} \tag{3.16}$$

pour des raisons pratiques pour effectuer une analyse sur la précision de l'estimateur $S_n(\tilde{f})$ défini plus loin (celle-ci est effectuée dans [18] et ne sera pas développée dans ce mémoire). Pour un échantillon de taille suffisamment grande, la différence entre l'estimateur classique et l'estimateur tronqué devient minime. Une illustration de l'estimation de la fonction de répartition d'une loi normale centrée réduite est disponible en Figure 3.4 pour des tailles d'échantillons n égales à 10, 50 et 200.

Ayant cette estimation de la fonction de répartition de la variable X_j, la fonction de transformation f_j peut être estimée par

$$\tilde{f}_j(x) \equiv \hat{\mu}_j + \hat{\sigma}_j \tilde{h}_j(x), \tag{3.17}$$

où $\tilde{h}_j(x) = \Phi^{-1}(\tilde{F}_j(x))$ et $\hat{\mu}_j$ et $\hat{\sigma}_j$ sont les moyenne et écart type (basé sur l'estimation biaisée de la variance) de l'échantillon :

$$\hat{\mu}_j \equiv \frac{1}{n}\sum_{i=1}^{n} x_{ij} \text{ et } \hat{\sigma}_j \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\hat{\mu}_j)^2}.$$

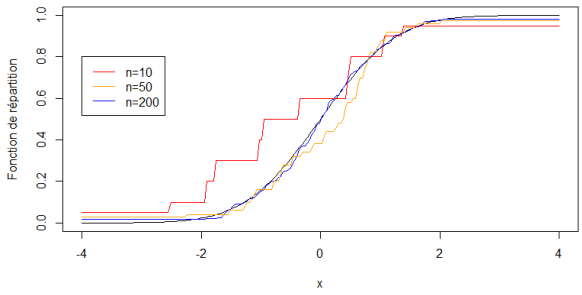

FIGURE 3.4 – Fonction de répartition d'une loi normale centrée réduite ainsi que son estimation tronquée pour des échantillons de taille $n = 10$, $n = 50$ et $n = 200$.

Illustration personnelle

Le même exemple que celui de la section précédente est utilisé, mais cette fois-ci, des données sont générées à partir des variables $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1, 5]$. La fonction de répartition de chaque variable sera alors estimée à l'aide de l'estimateur tronqué.

Tout d'abord, si les variables X_1 et X_2 sont indépendantes, alors la génération de données est simple. Voici en Figure 3.5, la représentation des vecteurs contenant les 200 données générées pour chaque variable selon leur situation sur la fonction de densité ainsi que leur situation dans le plan. La Figure 3.6 montre que la fonction de répartition estimée sur les données en utilisant l'estimateur tronqué est très proche de la fonction de répartition théorique des deux distributions. Le paramètre δ_n a été fixé comme défini en (3.16), ainsi $\delta_n \simeq 0.0163$ avec $n = 200$.

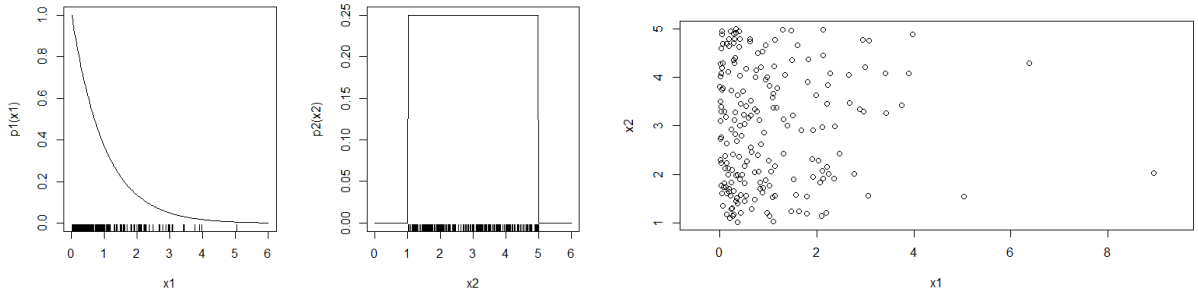

FIGURE 3.5 – Représentation des vecteurs de données générés pour les variables $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1, 5]$ selon leur situation sur la fonction de densité ainsi que dans le plan.

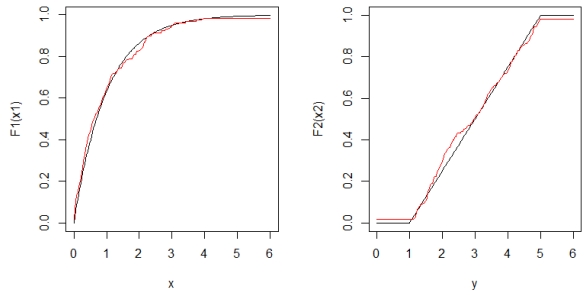

FIGURE 3.6 – Comparaison entre les fonctions de répartition théoriques (en noir) des variables $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1,5]$ et leur estimation tronquée basée sur les données générées (en rouge).

Il est maintenant possible de voir si la distribution jointe du vecteur $(f_1(X_1), f_2(X_2))$ est bien la distribution normale de vecteur moyen $(1,3)$ et dont la matrice de covariances est la matrice diagonale

$$\begin{pmatrix} 1 & 0 \\ 0 & \frac{4}{3} \end{pmatrix}. \tag{3.18}$$

Comme, en pratique, les distributions théoriques des variables sont inconnues, il est possible d'estimer les transformations f_j comme dans (3.17). Un nuage de points représentant le vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ est illustré en Figure 3.7.

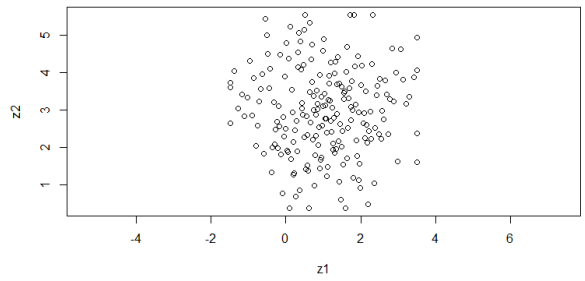

FIGURE 3.7 – Nuage de points représentant le vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$, $X_2 \sim \text{Unif}[1,5]$ et $X_1 \amalg X_2$.

Par cette transformation, la densité jointe du vecteur $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ devrait suivre une loi binormale. Cependant, le test de Mardia rejette la normalité de ce vecteur. La représentation de la densité jointe de $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ estimée à l'aide des noyaux est reprise en Figure 3.8. Il est possible de voir que les courbes de niveau forment des cercles

centrés en $(1, 3)$ qui ne sont pas réguliers. Augmenter le nombre de points générés ne permet pas d'obtenir un non rejet de la normalité suivant le test de Mardia.

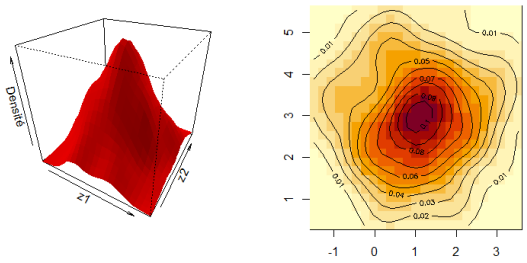

FIGURE 3.8 – Graphique et courbes de niveau de la densité jointe empirique du vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$, $X_2 \sim \text{Unif}[1, 5]$ et $X_1 \amalg X_2$.

Si, maintenant, les variables sont corrélées, alors, pour générer des données, il est possible d'utiliser la théorie des copules. Pour cet exemple, la copule normale est utilisée avec un paramètre de corrélation égal à $\rho = 0.1$. En Figure 3.9 se trouve le nuage de points représentant les 200 données générées.

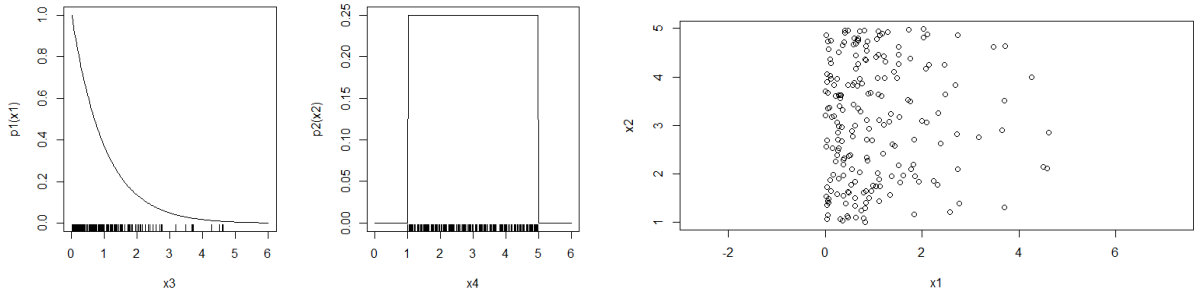

FIGURE 3.9 – Représentation des vecteurs de données générés pour les variables corrélées $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1, 5]$ selon leur situation sur la fonction de densité ainsi que dans le plan.

En effectuant les deux mêmes transformations que pour l'exemple non corrélé, le nuage de points représentant les variables normalisées peut être obtenu et se trouve en Figure 3.10. La densité du vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ estimée par noyau (Figure 3.11) est proche de la normale mais, à nouveau, les courbes de niveau ne sont par circulaires. Le test de Mardia confirme aussi le rejet de la normalité. Augmenter la taille de l'échantillon ne permet pas non plus d'obtenir des données dont la normalité pourrait être supposée.

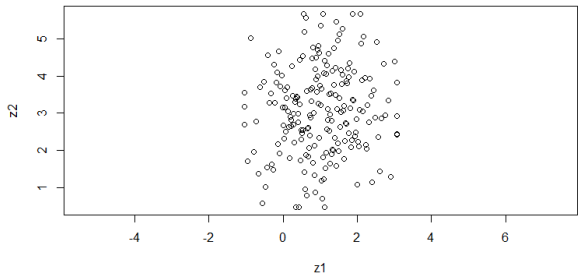

FIGURE 3.10 – Nuage de points représentant le vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1, 5]$.

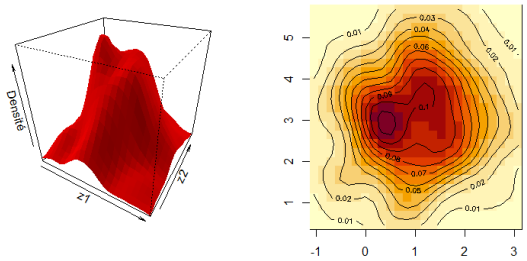

FIGURE 3.11 – Graphique et courbes de niveau de la densité jointe empirique du vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1, 5]$.

Une première conclusion laisse à penser que la technique de normalisation des données est vérifiée en théorie mais qu'en pratique, lorsque la fonction de répartition des données doit être estimée, elle n'aboutit pas au résultat demandé.

3.4 Simulations sur les tests de multinormalité

Comme, au vu de l'illustration précédente, le test de Mardia rejette régulièrement la normalité des données obtenues après la transformation, il peut être judicieux de s'interroger sur l'adéquation du test. C'est pourquoi, dans cette section, des simulations sont effectuées pour comparer deux tests de multinormalité. Ces tests sont le test de Mardia et celui de Henze-Zirkler. Ils sont tous deux disponibles dans R dans la librairie MVN.

Ces deux tests sont constitués comme ceci :

$$\begin{cases} H_0 : X \sim \mathcal{N}_d(\mu, \Theta) \\ H_1 : X \text{ ne suit pas une loi mulinormale.} \end{cases}$$

Étant donné qu'il a été montré dans ce chapitre que, après transformation, les données suivent bien une loi multinormale, elles sont donc sous l'hypothèse nulle. C'est pourquoi il est intéressant d'étudier la probabilité de rejeter H_0, alors que les données sont sous H_0 ; c'est-à-dire l'erreur de première espèce des tests. Celle-ci doit être proche de 0.05, l'erreur fixée avant d'effectuer les tests.

Le test de Mardia [13] est composé de deux statistiques de test : aplatissement et dissymétrie. Celles-ci sont définies par

$$b_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big((x_i - \bar{x})^T S^{-1} (x_j - \bar{x}) \Big)^3 \tag{3.19}$$

pour la dissymétrie et

$$b_2 = \frac{1}{n} \sum_{i=1}^{n} \Big((x_i - \bar{x})^T S^{-1} (x_i - \bar{x}) \Big)^2 \tag{3.20}$$

pour l'aplatisement. De plus, les distributions asymptotiques des deux statistiques de test sont la loi χ^2 pour $\frac{nb_1}{6}$ et la loi normale centrée et réduite pour $(b_2 - d(d+2))\sqrt{\frac{n}{8d(d+2)}}$. Le test de Henze-Zirkler est, quant à lui, défini par une statistique de test unique et plus difficile à comprendre, celle-ci est basée sur la fonction caractéristique empirique et est détaillé dans [32] .

Une première simulation consiste simplement à générer $N = 1000$ fois des données de taille n (qui varie) selon une loi binormale de vecteur moyen nul et de matrice de covariances la matrice identité. Le taux de rejet lors de ces 1000 générations de données est alors calculé pour obtenir le taux d'erreur de première espèce. Le graphique de la Figure 3.12 montre que, lorsque les données sont générées selon une loi multinormale, le taux d'erreur de première espèce du test de Mardia est plus élevé que celui attendu de 0.05. Le test de Henze-Zirkler, quant à lui, respecte mieux ce taux.

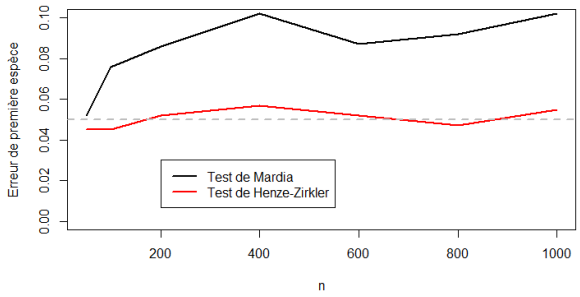

FIGURE 3.12 – Graphique comparant, en fonction de la taille de l'échantillon, le taux d'erreur de première espèce des tests de Mardia et de Henze-Zirkler lorsque les données sont générées selon une loi binormale de moyenne nulle et de matrice de covariances la matrice identité.

Au vu des valeurs du taux d'erreur de première espèce pour le test de Mardia, cela laisse à penser que, comme ce test comporte deux statistiques, une correction doit être appliquée. Lorsque la correction de Bonferroni est appliquée, l'erreur de première espèce du test de Mardia est située autour de 0.05 (Figure 3.13). Cette correction permet de corriger le seuil de significativité lors de tests multiples. L'erreur fixée à 0.05 des différents tests est divisée par le nombre de tests à effectuer. Dans le cas du test de Mardia, la p-valeur des deux tests sera comparée à l'erreur de 0.025.

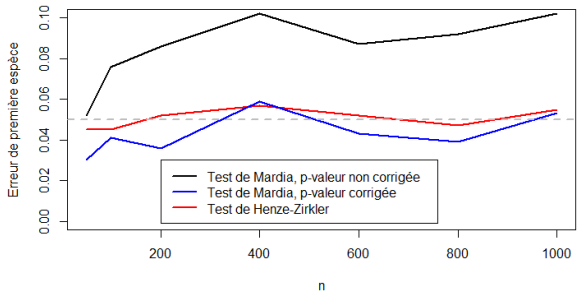

FIGURE 3.13 – Graphique comparant, en fonction de la taille de l'échantillon, le taux d'erreur de première espèce des tests de Mardia (avec et sans correction de Bonferroni) et de Henze-Zirkler lorsque les données sont générées selon une loi binormale de moyenne nulle et de matrice de covariances la matrice identité.

Une seconde simulation consiste à générer des données bivariées dont les 2 variables sont indépendantes et suivent une loi de probabilité connue. Dans ce cas, il a été démontré que, après transformation des données à l'aide des fonctions f_j de (3.6), les données suivent une loi binormale. Les fonctions de transformation sont donc connues et peuvent directement être utilisées. Pour ce faire, les deux lois étudiées lors de l'exemple de la section 3.2. sont utilisées. Comme précédemment (voir Figure 3.14), les tests de Henze-Zirkler et de Mardia, lorsque la correction de Bonferroni est appliquée, respectent le taux d'erreur de première espèce attendu.

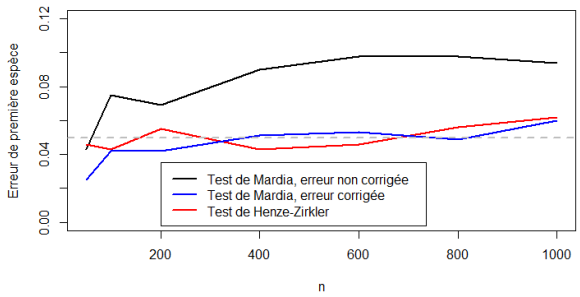

FIGURE 3.14 – Graphique comparant, en fonction de la taille de l'échantillon, l'erreur de première espèce des tests de Mardia (avec et sans correction de Bonferroni) et de Henze-Zirkler lorsque les données sont le vecteur aléatoire $(f_1(X_1), f_2(X_2))$ où $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}\,[1,5]$ et $X_1 \amalg X_2$.

La dernière simulation consiste à générer les mêmes données que pour la simulation précédente mais cette fois-ci, les fonctions f_j sont estimées comme dans (3.17). La multinormalité des données n'a été prouvée que dans le cas où les fonctions de transformation sont connues mais, en les estimant, les données transformées devraient tout de même être très proche d'une loi multinormale. Le résultat de cette simulation, repris en Figure 3.15, montre que, plus la taille de l'échantillon est grande, plus le test de Mardia (que la correction de Bonferroni soit appliquée ou non) rejettera la multinormalité. Cela semble plutôt étrange car, plus l'échantillon est grand, meilleure sera l'estimation de la fonction de transformation.

Il reste à déterminer si une des deux statistiques du test de Mardia mène davantage au rejet que l'autre. La Figure 3.16 montre que c'est la statistique d'aplatissement qui mène principalement au rejet de la normalité. Cette statistique devrait être distribuée selon la loi normale centrée et réduite. Or, au vu des histogrammes de la Figure 3.17, lorsque les fonctions de transformation sont estimées, la moyenne des statistiques de test se rapproche de -2. Cela signifie que la moyenne des carrés des carrés des distances de Mahalanobis est plus petite que la valeur moyenne théorique de $d(d+2)$.

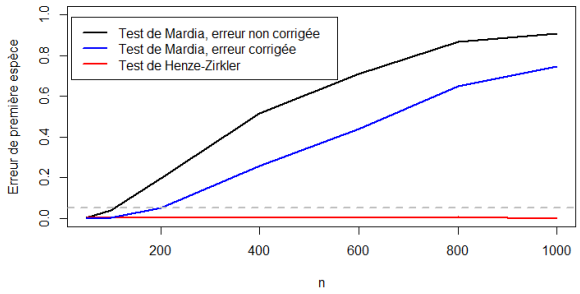

FIGURE 3.15 – Graphique comparant, en fonction de la taille de l'échantillon, l'erreur de première espèce des tests de Mardia (avec et sans correction de Bonferroni) et de Henze-Zirkler lorsque les données sont le vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1,5]$ et $X_1 \amalg X_2$.

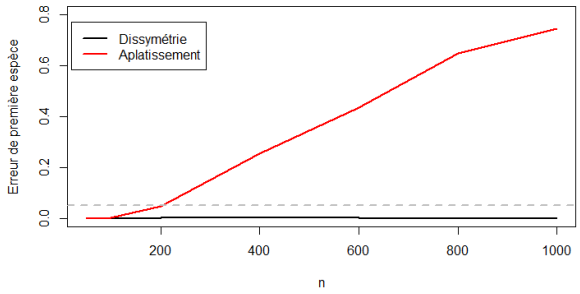

FIGURE 3.16 – Graphique comparant, en fonction de la taille de l'échantillon, l'erreur de première espèce des tests d'aplatissement et de dissymétrie lorsque les données sont le vecteur aléatoire $(\tilde{f}_1(X_1), \tilde{f}_2(X_2))$ où $X_1 \sim \text{Exp}(1)$ et $X_2 \sim \text{Unif}[1,5]$ et $X_1 \amalg X_2$.

En conclusion, il semblerait plus adéquat de se baser sur le test de Henze-Zirkler car la multinormalité des données a été prouvée théoriquement dans ce chapitre et qu'elle devrait être vérifiée également lors du passage aux estimations. Cette même constatation est effectuée lorsque la dimension des données est plus élevée. De plus, en se basant sur le test de Henze-Zirkler, la normalité des données obtenues lors des exemples de la section précédente n'est plus rejetée alors qu'elle l'était avec le test de Mardia.

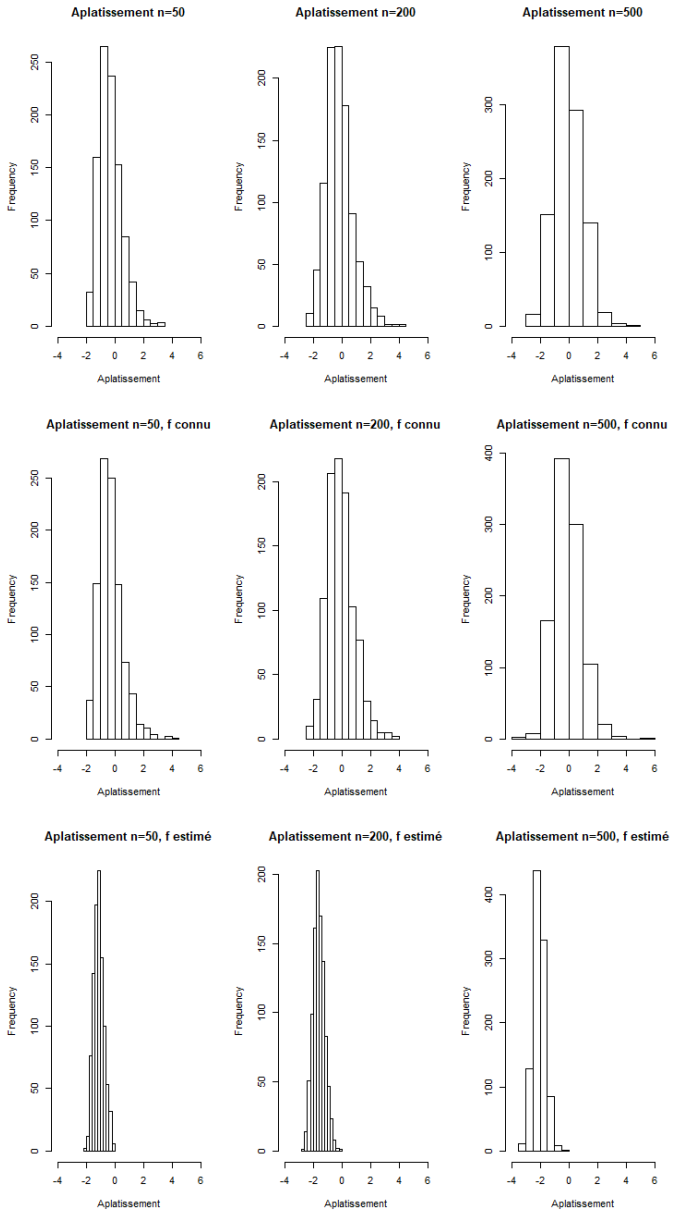

FIGURE 3.17 – Histogrammes représentant les valeurs de la statistique de test d'aplatissement, lors des 3 simulations effectuées ci-dessus, pour les tailles d'échantillon $n = 50$, 200 et 500.

3.5 Lien avec les modèles graphiques

Le fait que $Z \sim \mathcal{N}_d(\mu, \Sigma)$ implique que $Z_j \amalg Z_k | Z_{\setminus\{j,k\}}$ si et seulement si $\Theta_{jk} = (\Sigma^{-1})_{jk} = 0$. De plus, comme X_j est fonction de la variable Z_j uniquement, cela implique que $X_j \amalg X_k | X_{\setminus\{j,k\}}$ si et seulement si $\Theta_{jk} = 0$. Ce qui implique que la forme du graphe d'indépendance conditionnelle pour le non paranormal est reprise dans Θ, comme pour le cas normal paramétrique.

Plus généralement, le fait que la dépendance conditionnelle dépende des valeurs de Θ est vrai pour n'importe quel choix de restriction d'identification. Ainsi, il n'est pas nécessaire d'estimer μ et le vecteur $\sigma = (\sigma_1, \cdots, \sigma_d)$ pour estimer le graphe, comme le résultat suivant le montre :

Lemme 1. *Soit un vecteur aléatoire $X = (X_1, \ldots, X_d)$ tel que $X \sim NPN(\mu, \Sigma, f)$ où Σ est défini positif et $f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x))$. Soient*

$$h_j(x) = \Phi^{-1}(F_j(x)), \quad j = 1 \ldots, d \tag{3.21}$$

et Λ la matrice de covariance de $h(X) = (h_1(X_1), \cdots, h_d(X_d))$. Alors $X_j \amalg X_k | X_{\setminus\{j,k\}}$ si et seulement si $\Lambda^{-1}_{jk} = 0$.

Démonstration. La matrice de covariances Σ peut se réécrire sous la forme

$$\Sigma_{jk} = \mathrm{Cov}(Z_j, Z_k) = \mathrm{Cov}(\mu_j + \sigma_j h_j(X_j), \mu_k + \sigma_k h_k(X_k)) = \sigma_j \sigma_k \mathrm{Cov}(h_j(X_j), h_k(X_k)).$$

C'est pourquoi $\Sigma = D\Lambda D$ et $\Sigma^{-1} = D^{-1}\Lambda^{-1}D^{-1}$ où D est la matrice diagonale avec $\mathrm{diag}(D) = \sigma$. Il s'ensuit que $\Lambda^{-1}_{jk} = 0$ est équivalent à $\Sigma^{-1}_{jk} = 0$. □

En pratique, pour représenter le graphe de concentration du vecteur X, une fois que les fonctions f_j sont estimées, il ne reste plus qu'à estimer la matrice de concentration du vecteur aléatoire. Soit $S_n(\tilde{f})$ la matrice de covariances de $(\tilde{f}_1(X_{i1}), \ldots, \tilde{f}_d(X_{id}))$, $i = 1, \ldots, n$, où $\tilde{f}_j(X_{ij}) = (\tilde{f}_j(X_{1j}), \cdots, \tilde{f}_j(X_{n1}))$, $j = 1, \ldots, d$; c'est-à-dire

$$S_n(\tilde{f}) \equiv \frac{1}{n} \sum_{i=1}^{n} (\tilde{f}_j(x_{ij}) - \mu_n(\tilde{f}_j))(\tilde{f}_j(x_{ij}) - \mu_n(\tilde{f}_j))^T \quad \text{et}$$

$$\mu_n(\tilde{f}_j) \equiv \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_j(x_{ij}).$$

La matrice Θ peut être estimée en utilisant l'algorithme *graphical lasso* sur la matrice $S_n(\tilde{f})$, comme vu au chapitre 2.

Ainsi, une procédure en 2 étapes est utilisée pour estimer le graphe :

(1) Remplacer les observations, pour chaque variable, par leur transformation via les fonctions $\tilde{f}_j(x) = \Phi^{-1}(\tilde{F}_j(x))$, sujet à l'estimation tronquée.

(2) Appliquer le lasso aux données transformées pour estimer le graphe non dirigé.

3.5.1 Modèles graphiques du type *non paranormal* des données socio-économiques de la Wallonie

Pour effectuer le graphe représentant la base de données à la manière non paranormale, la procédure décrite dans la section 3.1.3 a été suivie. Une première remarque à faire est que l'hypothèse de normalité des données transformées est rejetée selon le test de Mardia, ce qui semble normal au vu des conclusions obtenues sur les simulations. Elle l'est aussi pour le test de Henze-Zirkler, car une des 19 variables ne suit pas une loi normale univariée. En effet, au vu des histogrammes de la Figure 3.18, la variable Elec ne suit pas une loi normale univariée après transformation. Cela est du au fait qu'aucune donnée ne se situe entre 200 et 300. Cette absence de données provoque un 'trou' dans l'histogramme après normalisation au niveau des valeurs de -50 à -20. Ce cas particulier de variables possédant un certain nombre de données très éloignées de la moyenne ne peut pas être réglé à l'aide de la transformation non paramétrique.

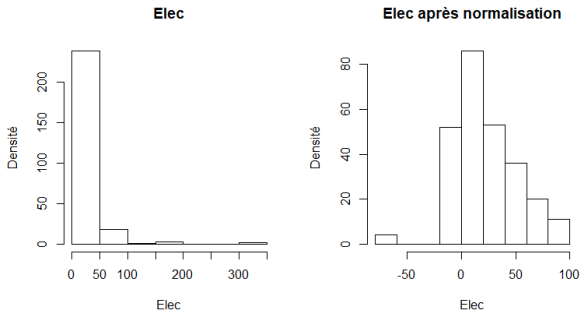

FIGURE 3.18 – Histogrammes représentant la variable Elec avant (à gauche) et après (à droite) normalisation.

La deuxième remarque est que la matrice de corrélation des données modifiées selon $\tilde{f}_j(x) \equiv \tilde{h}_j(x)$ donnée à la Figure 3.19 est fort similaire à celle pour la base de données de départ.

La dernière remarque est que, comme précédemment, le critère BIC est strictement croissant en fonction du paramètre de pénalisation λ (Figure 3.20). Il en est de même pour le critère AIC. Il n'est donc pas possible de ressortir de l'information en utilisant ces 2 critères. Comme dans le chapitre précédent, les valeurs de λ sont choisies pour que le graphe possède 36 et 18 arêtes. Ainsi au vu de la Figure 3.21 représentant la variation du nombre d'arêtes dans le graphe en fonction de la valeur du paramètre de pénalisation, ces nombres d'arêtes sont atteints pour $\lambda = 0.43$ et $\lambda = 0.62$ respectivement. Ces deux graphes sont repris en Figure 3.22.

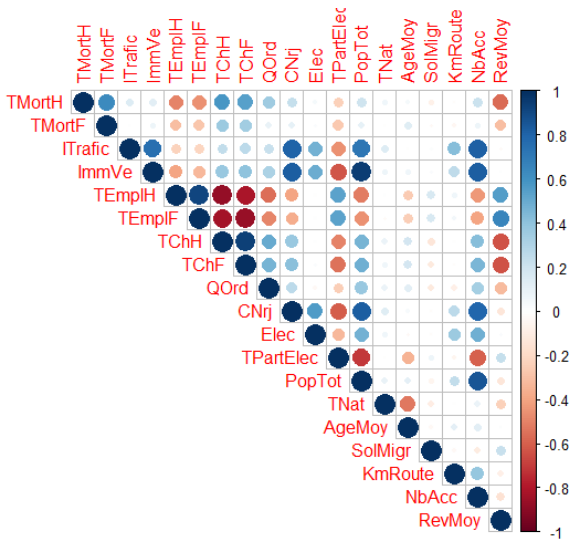

FIGURE 3.19 – Représentation de la matrice de corrélation de la base de données modifiée.

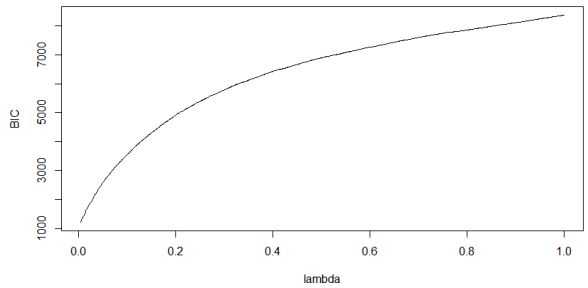

FIGURE 3.20 – Variation du BIC en fonction du paramètre de pénalisation λ.

Ces deux graphes sont relativement différents par rapport aux deux graphes du chapitre précédent basés sur la base de données de départ. Cela peut s'expliquer par le fait que les données normalisées ont été centrées et réduites. En effet, les fonctions de transformation utilisées sont les fonctions $\tilde{f}_j(x) = \Phi^{-1}(\tilde{F}_j(x))$ où les moyennes et variances marginales ne sont pas estimées. Ainsi, les graphes de cette section sont similaires aux graphes du chapitre précédent, basés sur les données dont les variables ont été centrées et réduites. Il est tout de même judicieux, dans le cas de cette base de données, de considérer les données réduites ou de baser l'analyse sur la matrice de corrélation puisque les échelles des diverses variables sont très différentes.

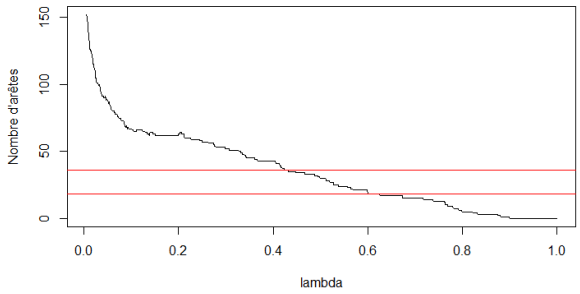

FIGURE 3.21 – Variation du nombre d'arêtes dans le graphe représentant la base de données, en fonction de la valeur de lambda.

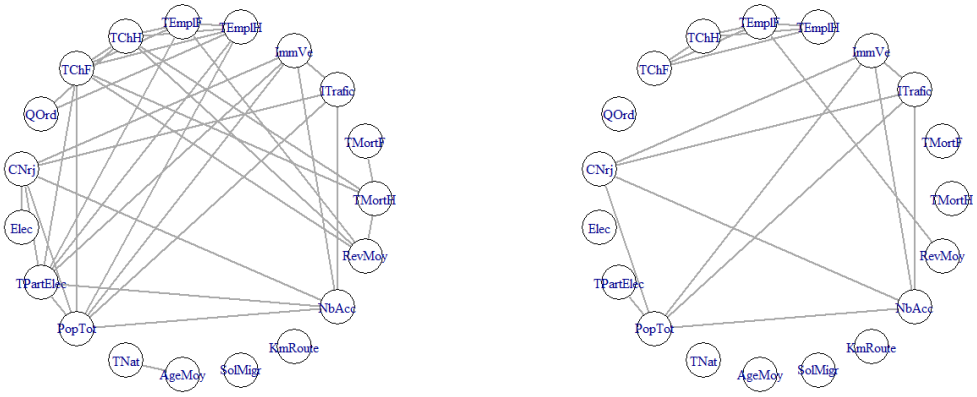

FIGURE 3.22 – Modèles graphique pour la base de données, créés en utilisant la normalisation non paranormale, contenant 36 arêtes (à gauche) et 18 arêtes (à droite).

La différence entre les deux techniques n'est pas très marquée. Cependant, sur les 18 arêtes du graphe, deux arêtes ont changé de place. Pour la technique *glasso*, les arêtes entre les variables TChH et RevMoy ainsi qu'entre TChF et RevMoy sont présentes. Dans le graphe créé à partir de la technique de normalisation non paranormale, ces arêtes sont situées entre les variables TPartElec et PopTot, ainsi qu'entre CNrj et ImmVe. Ce changement pourrait être expliqué par le fait que les dépendances entre les variables TPartElec et PopTot, ainsi que CNrj et ImmVe sont plus fortes que celles entre les variables TChH et RevMoy ainsi que TChF et RevMoy, mais que les premières sont des dépendances qui ne sont pas suffisamment linéaires que pour être reconnues par la technique *glasso*.

Chapitre 4

Modèles graphiques sous forme de forêt

Dans ce chapitre, une deuxième approche pour élargir les modèles graphiques est proposée, basée sur les articles [18] et [22]. Celle-ci consiste à imposer une structure au graphe. Différents types de structure peuvent être imposés. Dans ce mémoire, seule la structure d'arbre, et plus particulièrement de forêt, sera étudiée.

Cette méthode consiste à trouver la forêt qui représente le mieux la distribution des données. Ainsi, chaque paire de sommets est connectée par au plus un chemin. Le graphe créé possèdera donc très peu d'arêtes. Cette méthode peut s'effectuer par une procédure en trois étapes :

1. Estimer la fonction de densité p^* des données si celle-ci est inconnue ;
2. Imposer une structure de forêt au graphe final ;
3. Trouver la forêt et la densité qui peut être supportée par la forêt en question de manière à minimiser la distance de Kullback-Leibler entre cette densité et la densité p^* de départ.

Lorsque la fonction de densité est connue, la procédure débute directement à la deuxième étape. Lorsque celle-ci n'est pas connue, elle doit être estimée. Dans ce travail, ce sont des outils non paramétriques qui seront exploités. Ainsi, comme en pratique, la densité des données est rarement connue, la technique d'estimation de modèle graphique sous forme d'arbre est considérée comme étant une technique d'estimation non paramétrique. Dans ce chapitre, la matrice de concentration des données ne fera plus du tout l'objet de l'estimation de la structure du graphe, contrairement aux chapitres précédents.

4.1 Définition des outils

Soit p^* une densité de probabilité par rapport à la mesure de Lebesgue $\mu(\cdot)$ sur $\mathbb{R}^d$. Soit χ_j le domaine de définition des valeurs possibles de la j-ème variable et soit $\chi = \chi_1 \times ... \times \chi_d$.

Voici, pour commencer, quelques définitions utiles pour la construction d'une forêt.

Un graphe est une forêt s'il est acyclique. Si F est une forêt non dirigée possédant d sommets, avec l'ensemble de sommets $V_F = \{1, \ldots, d\}$ et l'ensemble d'arêtes $E_F \subset \{1, \ldots, d\} \times \{1, \ldots, d\}$, alors le nombre d'arêtes satisfait $|E_F| < d$.

Définition 5. Une fonction de densité $p(x)$ est *supportée par une forêt* F si elle peut être écrite comme

$$p_F(x) = \prod_{(i,j)\in E_F} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)} \prod_{k\in V_F} p_k(x_k), \tag{4.1}$$

où chaque $p_{ij}(x_i, x_j)$ est une densité bivariée sur $\chi_i \times \chi_j$, et chaque $p_k(x_k)$ est une densité univariée sur χ_k, avec $p_k(x_k) > 0$ pour tout $k \in V_F$.

Définition 6. La divergence de Kullback-Leibler $D(p||q)$ entre deux densités p et q est

$$D(p||q) = \int_\chi p(x) \ln \frac{p(x)}{q(x)} dx, \tag{4.2}$$

avec la convention que $0 \ln(0/q) = 0$, et $p \ln(p/0) = \infty$ pour $p \neq 0$.

Définition 7. Pour toute densité q, le risque négatif de la log-vraisemblance par rapport à une densité p est défini par

$$R(q) = -\mathbb{E}[\ln q(X)] = -\int_\chi p(x) \ln q(x) dx. \tag{4.3}$$

Illustration personnelle

Afin d'illustrer la définition 5, voici un exemple de fonction de densité supportée par une forêt. Soit F la forêt représentée en Figure 4.1.

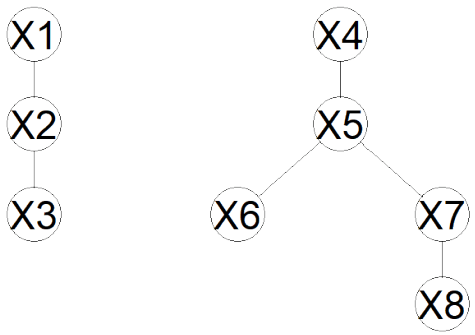

FIGURE 4.1 – Exemple de forêt possédant 8 sommets.

Toute fonction de densité supportée par cette forêt doit être de la forme

$$p_F(x) = \frac{p_{12}(x_1, x_2)p_{23}(x_2, x_3)p_{45}(x_4, x_5)p_{56}(x_5, x_6)p_{57}(x_5, x_7)p_{78}(x_7, x_8)}{p_2(x_2)p_5(x_5)p_5(x_5)p_7(x_7)}. \tag{4.4}$$

Dans cet exemple, les densités marginales bivariées sont choisies comme étant une fonction de densité binormale de vecteur moyen le vecteur nul et de matrice de covariances la matrice

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Elles ont donc la forme

$$p_{ij}(x_i, x_j) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2} \frac{1}{1-\rho^2} (x_i^2 + x_j^2 - 2\rho x_i x_j) \right).$$

Les densités univariées ont alors la forme

$$p_i(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2} x_i^2 \right).$$

La fonction de densité jointe est obtenue en effectuant le produit de (4.4) :

$$p_F(x) = \frac{1}{(2\pi)^4} \left(\frac{1}{\sqrt{1-\rho^2}} \right)^6 \exp\left(\frac{-1}{2} \frac{1}{1-\rho^2} (x_1^2 + \rho^2 x_2^2 + x_3^2 + x_4^2 + (2\rho^2 - 1)x_5^5 + x_6^2 + \rho^2 x_7^2 + x_8^2 \right.$$
$$\left. - 2\rho(x_1 x_2 + x_2 x_3 + x_4 x_5 + x_5 x_6 + x_5 x_7 + x_7 x_8) \right).$$

Cette fonction est la fonction de densité d'une loi multinormale dont le vecteur moyen est le vecteur nul et dont la matrice de concentration est

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 & 0 & 0 & 0 \\ -\rho & \rho^2 & -\rho & 0 & 0 & 0 & 0 & 0 \\ 0 & -\rho & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -\rho & 0 & 0 & 0 \\ 0 & 0 & 0 & -\rho & 2\rho^2 - 1 & -\rho & -\rho & 0 \\ 0 & 0 & 0 & 0 & -\rho & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\rho & 0 & \rho^2 & -\rho \\ 0 & 0 & 0 & 0 & 0 & 0 & -\rho & 1 \end{pmatrix}.$$

4.2 Premier cas : la densité jointe est connue

Il est maintenant utile d'exploiter ces définitions dans les différentes étapes de la construction d'un graphe. Lorsque la densité jointe des données p^* est connue, l'étape d'estimation n'est pas utile. Il reste donc à trouver une forêt et une densité supportée par cette forêt qui seront les plus représentatifs de la densité p^*.

Soit $\mathcal{F}_d$ la famille de forêts avec d sommets, et soit $\mathcal{P}_d$ la famille correspondante de densités

$$\mathcal{P}_d = \{p \geq 0 : \int_\chi p(x)d\mu(x) = 1, \text{ et } p(x) \text{ satisfait (4.1) pour un } F \in \mathcal{F}_d\}. \tag{4.5}$$

La densité qui peut être supportée par une forêt qui représente le mieux les données est choisie de telle sorte qu'elle soit la plus proche possible de la densité p^*. C'est-à-dire qu'elle sera choisie de telle sorte à minimiser la distance de Kullback-Leibler et à appartenir à l'ensemble $\mathcal{P}_d$ des densités supportées par une forêt à d sommets. Elle peut être définie par

$$q^* = \arg\min_{q \in \mathcal{P}_d} D(p^*||q). \tag{4.6}$$

Cette densité q^* est supportée par une forêt, elle peut donc s'écrire comme en (4.1). De plus les densités univariées et bivariées utilisées dans cette formule sont les densités marginales univariées et bivariées de p^*. Cela est démontré par la proposition suivante dont la preuve a été rédigée en référence à l'article [2].

Proposition 4. *Soit q^* défini comme en (4.6). Il existe une forêt $F^* \in \mathcal{F}_d$ telle que*

$$q^* = p^*_{F^*} = \prod_{(i,j)\in E_{F^*}} \frac{p^*_{ij}(x_i, x_j)}{p^*_i(x_i)p^*_j(x_j)} \prod_{k \in V_{F^*}} p^*_k(x_k), \tag{4.7}$$

où $p^*_{ij}(x_i, x_j)$ et $p^*_k(x_k)$ sont les densités marginales bivariées et univariées de p^*.

Démonstration. Soit q une distribution de $\mathcal{P}_d$ qui est telle que

$$q(x) = \prod_{(i,j)\in E_{F^*}} \frac{q_{ij}(x_i, x_j)}{q_i(x_i)q_j(x_j)} \prod_{k \in V_{F^*}} q_k(x_k),$$

pour un certain F^*.

Comme $q(x)$ est un produit de fonctions de paire de variables liées par une arête et de fonctions de variable unique, et comme $p^*(x)$ et $p^*_{F^*}(x)$ ont les mêmes distributions marginales sur les cliques du graphe, c'est-à-dire les mêmes distributions marginales univariées et bivariées, il en ressort que

$$\mathbb{E}_{p^*}[\ln q(X)] = \mathbb{E}_{p^*_{F^*}}[\ln q(X)],$$

où $\mathbb{E}_p[\ln q(X)] = \int_\chi p(x) \ln q(x) dx$.

En appliquant ce même raisonnement à $p^*_{F^*}(x)$, cela donne

$$\mathbb{E}_{p^*}[\ln p^*_{F^*}(X)] = \mathbb{E}_{p^*_{F^*}}[\ln p^*_{F^*}(X)].$$

Ces deux égalités permettent de prouver l'identité de Pythagore :

$$\begin{aligned}
D(p^*||q) &= \int_\chi p^*(x) \ln \frac{p^*(x)}{q(x)} dx \\
&= \int_\chi p^*(x) \ln p^*(x) dx - \int_\chi p^*(x) \ln q(x) dx \\
&= \mathbb{E}_{p^*}[\ln p^*(X)] - \mathbb{E}_{p^*}[\ln q(X)] \\
&= \mathbb{E}_{p^*}[\ln p^*(X)] - \mathbb{E}_{p^*_{F^*}}[\ln q(X)] \\
&= \mathbb{E}_{p^*}[\ln p^*(X)] - \mathbb{E}_{p^*_{F^*}}[\ln p^*_{F^*}(X)] + \mathbb{E}_{p^*_{F^*}}[\ln p^*_{F^*}(X)] - \mathbb{E}_{p^*_{F^*}}[\ln q(X)] \\
&= \mathbb{E}_{p^*}[\ln p^*(X)] - \mathbb{E}_{p^*}[\ln p^*_{F^*}(X)] + \mathbb{E}_{p^*_{F^*}}[\ln p^*_{F^*}(X)] - \mathbb{E}_{p^*_{F^*}}[\ln q(X)] \\
&= D(p^*||p^*_{F^*}) + D(p^*_{F^*}||q).
\end{aligned}$$

Cela montre que la distribution q qui minimise $D(p^*||q)$ est donnée par $p^*_{F^*}$, car la distance de Kullback-Leibler est toujours positive ou nulle. □

Minimiser la distance de Kullback-Leibler revient à minimiser le risque négatif de la log-vraisemblance. En effet, il est démontré par la proposition suivante que la densité q^* définie en (4.6) minimise le risque négatif de la log-vraisemblance par rapport à la densité p^*.

Proposition 5. *Soit q^* défini par $q^* = \underset{q \in \mathcal{P}_d}{arg\ min}\ D(p^*||q)$, alors*

$$q^* = \underset{q \in \mathcal{P}_d}{arg\ min}\ R(q). \tag{4.8}$$

Démonstration.

$$\begin{aligned}
q^* &= \underset{q \in \mathcal{P}_d}{\arg\min}\ D(p^*||q) \\
&= \underset{q \in \mathcal{P}_d}{\arg\min}\ \int_\chi p^*(x) \ln \left(\frac{p^*(x)}{q(x)} \right) dx \\
&= \underset{q \in \mathcal{P}_d}{\arg\min}\ \left(\int_\chi p^*(x) \ln(p^*(x)) dx - \int_\chi p^*(x) \ln(q(x)) dx \right) \\
&= \underset{q \in \mathcal{P}_d}{\arg\min}\ - \int_\chi p^*(x) \ln(q(x)) dx \\
&= \underset{q \in \mathcal{P}_d}{\arg\min}\ R(q).
\end{aligned}$$

□

Pour trouver la forêt représentant au mieux la densité, il suffit d'étudier l'information mutuelle entre les variables.

Définition 8. Soit (X_i, X_j) un couple de variables de densité de probabilité jointe $p_{ij}(x_i, x_j)$ et dont les densités marginales sont données par $p_i(x_i)$ et $p_j(x_j)$, alors l'information mutuelle entre des variables X_i et X_j est définie par

$$I(X_i, X_j) = \int_{\chi_i} \int_{\chi_j} p_{ij}(x_i, x_j) \ln \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)} dx_i dx_j. \tag{4.9}$$

Définition 9. Soit X_k une variable de densité de probabilité $p_k(x_k)$, alors l'entropie de cette variable est définie par

$$H(X_k) = -\int_{\chi_k} p_k(x_k) \ln p_k(x_k) dx_k.$$

Proposition 6. *La forêt F^* représentant au mieux la densité p^* est donnée par*

$$F^* = arg \max_F \sum_{(i,j) \in E_F} I(X_i, X_j).$$

Démonstration. Le risque minimum est défini par $R^* = R(q^*)$. En utilisant la proposition 4,

$$\begin{aligned} R^* = R(q^*) &= R(p^*_{F^*}) && (4.10)\\ &= -\int_{\chi} p^*(x) \left(\sum_{(i,j) \in E_{F^*}} \ln \frac{p^*_{ij}(x_i, x_j)}{p^*_i(x_i) p^*_j(x_j)} + \sum_{k \in V_{F^*}} \ln p^*_k(x_k) \right) dx && (4.11)\\ &= -\sum_{(i,j) \in E_{F^*}} I(X_i; X_j) + \sum_{k \in V_{F^*}} H(X_k). && (4.12) \end{aligned}$$

Ainsi, la forêt optimale peut être trouvée en minimisant (4.12). Comme le terme d'entropie $H(X) = \sum_k H(X_k)$ est constant sur toutes les forêts, cela se résume à minimiser le terme d'information mutuelle. □

Pour trouver la forêt optimale, il suffit donc de trouver une forêt de poids maximum couvrant le graphe pondéré, où le poids de l'arête connectant les sommets i et j est $I(X_i; X_j)$. L'algorithme de Kruskal est un algorithme glouton qui garantit de trouver un arbre couvrant de poids maximum d'un graphe pondéré. Cet algorithme est détaillé dans [16] et mentionné dans les notes de cours de "Théorie des graphes" de M. Rigo. A chaque étape, l'algorithme ajoute une arête connectant la paire de variables qui a une information mutuelle maximale parmi toutes les paires qui ne sont pas encore visibles par l'algorithme, si faire cela ne crée pas de boucle. Quand l'algorithme s'arrête plus tôt, après que $k < d - 1$ arêtes aient été ajoutées, cela mène à la meilleure forêt pondérée possédant k arêtes.

Voici les détails de l'algorithme de Kruskal tel qu'utilisé dans le cas de l'estimation d'une densité sous forme de forêt.

Algorithme 3. ***Algorithme de Kruskal pour construire un arbre de poids maximal.***

Données : La base de données.

Initialisation : Calculer $I(X_i, X_j)$ *pour tout* $i \neq j$.

Poser $E^{(0)} = \emptyset$.

Pour $k = 1, \cdots, d-1$;

(1) Poser $(i^{(k)}, j^{(k)}) \leftarrow$ *arg max*$_{(i,j)} I(X_i, X_j)$ *tel que* $E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$ *ne contient pas de cycle ;*

(2) $E^{(k)} \leftarrow E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$.

Sortie : l'arbre $\hat{F}^{(d-1)}$ *avec l'ensemble d'arêtes* $E^{(d-1)}$. *L'algorithme s'arrête plus tôt si* $I(X_{i^{(k)}}, X_{j^{(k)}})$ *est nul et cette arête n'est pas insérée au graphe.*

Illustration personnelle

Pour illustrer la technique d'estimation du graphe sous forme de forêt lorsque la fonction de densité p^* est connue, il suffit simplement de pouvoir calculer l'information mutuelle entre chaque paire de variables pour pouvoir appliquer ensuite l'algorithme de Kruskal. Pour pouvoir calculer aisément la matrice d'information mutuelle, la densité p^* est choisie comme étant la fonction de densité d'une loi normale multivariée. En effet, dans ce cas, l'information mutuelle entre chaque paire de variables (X, Y) est donnée par

$$\frac{-1}{2} \ln(1 - \rho^2), \tag{4.13}$$

où ρ est le coefficient de corrélation entre ces deux variables. Ce résultat demande une étude plus approfondie de l'entropie de vecteurs gaussiens et est démontré dans le chapitre 7 de [17].

La moyenne de chaque variable n'influence pas les valeurs d'information mutuelle ; seule la matrice de covariances est utile pour estimer le graphe. Celle-ci est choisie comme étant la matrice par bloc suivante

$$\Sigma = \begin{pmatrix} 1 & 0.7 & 0.7 & 0.7 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 2 & 0.7 & 0.7 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 3 & 0.7 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 0.7 & 4 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 0.7 & 0.7 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 4 & 0.4 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.4 & 3 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.4 & 0.4 & 2 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 1 \end{pmatrix}.$$

Comme la variance n'est pas identique pour les 10 variables, la matrice de corrélation n'est pas identique à Σ. L'information mutuelle entre chaque paire de variables sera donc différente au sein de chaque bloc et entre les 2 blocs. La matrice d'information mutuelle, obtenue à l'aide de la formule (4.13), est alors

$$\begin{pmatrix} 0 & 0.1405 & 0.0892 & 0.0653 & 0.0516 & 0 & 0 & 0 & 0 & 0 \\ 0.1405 & 0 & 0.0426 & 0.0316 & 0.0251 & 0 & 0 & 0 & 0 & 0 \\ 0.0892 & 0.0426 & 0 & 0.0208 & 0.0166 & 0 & 0 & 0 & 0 & 0 \\ 0.0653 & 0.0316 & 0.0208 & 0 & 0.0124 & 0 & 0 & 0 & 0 & 0 \\ 0.0516 & 0.0251 & 0.0166 & 0.0124 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0040 & 0.0054 & 0.0081 & 0.0163 \\ 0 & 0 & 0 & 0 & 0 & 0.0040 & 0 & 0.0067 & 0.0101 & 0.0204 \\ 0 & 0 & 0 & 0 & 0 & 0.0054 & 0.0067 & 0 & 0.0135 & 0.0274 \\ 0 & 0 & 0 & 0 & 0 & 0.0081 & 0.0101 & 0.0135 & 0 & 0.0417 \\ 0 & 0 & 0 & 0 & 0 & 0.0163 & 0.0204 & 0.0274 & 0.0417 & 0 \end{pmatrix}. \tag{4.14}$$

Il reste à appliquer l'algorithme de Kruskal. La première arête ajoutée au graphe est celle entre les variables X_1 et X_2 car $I(X_1, X_2) = 0.1405$ est la valeur la plus élevée de la matrice. Ensuite, les arêtes entre X_1 et X_3, entre X_1 et X_4 et entre X_1 et X_5 sont ajoutées. La prochaine valeur la plus élevée dans la matrice est $I(X_2, X_3)$, or l'arête entre X_2 et X_3 ne peut être ajoutée car sinon un cycle serait formé. Il en est de même pour toutes les arêtes connectant des variables du premier bloc. Ainsi, les prochaines arêtes ajoutées doivent relier uniquement des variables du deuxième bloc. Les dernières arêtes ajoutées sont alors celles entre les variables X_9 et X_{10}, X_8 et X_{10}, X_7 et X_{10} ainsi que X_6 et X_{10}. Le graphe possède alors 8 arêtes. Il pourrait néanmoins en posséder 9 tout en restant un arbre mais les seules arêtes qu'il est encore possible d'ajouter sans former un cycle sont des arêtes liant une variable du bloc 1 à une variable du bloc 2 ; or, pour ces couples de variables, l'information mutuelle est nulle. Le graphe estimé sur base de cette fonction de densité possèdera donc 8 arêtes ; celui-ci est représenté en Figure 4.2.

4.3 Deuxième cas : la densité jointe est inconnue

La procédure ci-dessus n'est applicable en pratique que si la vraie densité p^* est connue. Lorsque ce n'est pas le cas, l'information mutuelle $I(X_i; X_j)$ en (4.9) doit être estimée. Soit $(X_{ij})_{1\leq i\leq n, 1\leq j\leq d}$ un échantillon aléatoire indépendant et identiquement distribué selon la loi p^* de $\mathbb{R}^d$. La $j^{\text{ème}}$ variable sera notée $X_j = (X_{1j}, \ldots, X_{nj})$. La façon la plus naturelle d'estimer $I(X_i; X_j)$ consiste à prendre comme estimation $\hat{I}_n(X_i; X_j)$ donnée par

$$\hat{I}_n(X_i; X_j) = \int_{\chi_i \times \chi_j} \hat{p}_{ij,n}(x_i, x_j) \ln \frac{\hat{p}_{ij,n}(x_i, x_j)}{\hat{p}_{i,n}(x_i)\hat{p}_{j,n}(x_j)} dx_i dx_j, \tag{4.15}$$

où $\hat{p}_{ij,n}(x_i, x_j)$ et $\hat{p}_{i,n}(x_i)$ sont les densités bivariées et univariées estimées. Plusieurs techniques existent pour estimer des densités. Dans ce mémoire, en suivant l'article [18], la

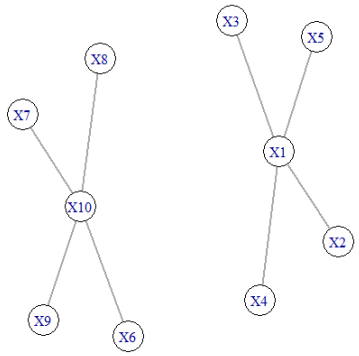

FIGURE 4.2 – Modèle graphique sous forme de forêt représentant la fonction de densité normale dont la matrice de covariances est Σ.

technique d'estimation par noyau sera utilisée.

En ayant cette approximation de l'information mutuelle $\hat{M}_n = [\hat{I}_n(X_i; X_j)]$, tout comme dans le cas précédent, l'algorithme de Kruskal peut être appliqué pour trouver la meilleure structure pour $\hat{F}_n$.

Le nombre d'arêtes de $\hat{F}_n$ contrôle le nombre de densités marginales estimées qui font effectivement partie de l'estimateur final de la densité, il est donc intéressant de se demander si ce nombre ne pourrait pas être inférieur à $d-1$, nombre maximal d'arêtes de l'arbre. C'est pourquoi la procédure en 2 étapes suivante est utilisée. Premièrement, les données sont séparées aléatoirement en deux ensembles de données. Les ensembles $\mathcal{D}_1$ et $\mathcal{D}_2$ contiennent les indices des observations comprises dans chaque groupe et sont de tailles n_1 et n_2 respectivement, pour des tailles à spécifier (voir section 4.4). Ensuite les étapes suivantes sont appliquées, où la notation F_n^k fait référence à la forêt F possédant k arêtes et créée à partir de n observations :

(1) En utilisant $\mathcal{D}_1$, construire l'estimation par noyau de la densité des marginales univariées et bivariées et calculer $\hat{I}_{n_1}(X_i; X_j)$ pour $i, j \in \{1, \cdots, d\}$ avec $i \neq j$. Construire un arbre complet $\hat{F}_{n_1}^{(d-1)}$, en utilisant l'algorithme de Kruskal.

(2) En utilisant $\mathcal{D}_2$, enlever des arêtes à l'arbre $\hat{F}_{n_1}^{(d-1)}$ pour trouver une forêt $\hat{F}_{n_1}^{\hat{k}}$ avec $\hat{k}$ arêtes, pour $0 \leq \hat{k} \leq d-1$.

Une fois que $\hat{F}_{n_1}^{\hat{k}}$ est obtenu à l'étape (2), $\hat{p}_{\hat{F}_{n_1}^{\hat{k}}}$ peut être calculé à l'aide de (4.1) en utilisant les estimations par noyau des densités construites à l'étape 1.

Voici le détail des deux étapes.

Étape 1 : Construire une séquence de forêts

Cette étape est basée sur l'ensemble de données $\mathcal{D}_1$. Soit $K(\cdot)$ une fonction de noyau univariée. En ayant un point d'évaluation (x_i, x_j), une façon d'estimer une densité bivariée par noyau est d'utiliser un noyau multiplicatif, c'est à dire que la densité bivariée sera estimée à l'aide d'un produit de deux noyaux univariés. Ainsi, la densité bivariée estimée par noyau pour (X_i, X_j) basée sur les observations $\{x_{si}, x_{sj}\}_{s\in\mathcal{D}_1}$ est définie par

$$\hat{p}_{ij,n_1}(x_i, x_j) = \frac{1}{n_1}\sum_{s\in\mathcal{D}_1}\frac{1}{h_{2i}h_{2j}}K\Big(\frac{x_{si}-x_i}{h_{2i}}\Big)K\Big(\frac{x_{sj}-x_j}{h_{2j}}\Big), \tag{4.16}$$

où $h_{2i} > 0$ et $h_{2j} > 0$ sont les paramètres de fenêtre bivariés dépendant des variables X_i et X_j prises en compte. La densité univariée estimée par noyau $\hat{p}_{k,n_1}(x_k)$ pour X_k est

$$\hat{p}_{k,n_1}(x_k) = \frac{1}{n_1}\sum_{s\in\mathcal{D}_1}\frac{1}{h_{1k}}K\Big(\frac{x_{sk}-x_k}{h_{1k}}\Big), \tag{4.17}$$

où $h_{1k} > 0$ est le paramètre de fenêtre univarié dépendant de la variable X_k. Davantage de détails sur la technique d'estimation par noyau se trouvent dans le cours intitulé "Statistique non paramétrique" de G. Haesbroeck.

Lorsque ces deux densités sont estimées, l'information mutuelle $\hat{I}_{n_1}(X_i, X_j)$ peut être calculée pour tout $i \neq j$. Une fois que la matrice $d \times d$ d'information mutuelle $\hat{M}_{n_1} = [\hat{I}_{n_1}(X_i; X_j)]$ est obtenue, l'algorithme de Kruskal peut être appliqué pour trouver un arbre de poids maximal. Celui-ci est le même que précédemment, seules les notations sont adaptées.

Algorithme 4. ***Algorithme de Kruskal pour construire un arbre de poids maximal.***

Données : La base de données contenant les observations reprises dans $\mathcal{D}_1$ ainsi que la taille n_1 de cette classe.

Initialisation : Calculer $\hat{M}_{n_1}$ grâce à (4.16) *et* (4.17).

Poser $E^{(0)} = \emptyset$.

Pour $k = 1, \cdots, d-1$;

(1) Poser $(i^{(k)}, j^{(k)}) \leftarrow$ arg $max_{(i,j)}\hat{M}_{n_1}(i,j)$ tel que $E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$ ne contient pas de cycle ;

(2) $E^{(k)} \leftarrow E^{(k-1)} \cup \{(i^{(k)}, j^{(k)})\}$.

Sortie : l'arbre $\hat{F}_{n_1}^{(d-1)}$ avec l'ensemble d'arêtes $E^{(d-1)}$.

Étape 2 : Choisir une taille de forêt

L'arbre complet $\hat{F}_{n_1}^{(d-1)}$ obtenu à l'étape 1 peut avoir une variance élevée quand la dimension d est grande, ce qui amène à de l'*overfitting* dans l'estimation de la densité. En effet, l'information mutuelle pour les dernières arêtes ajoutées au graphe devient de plus en plus petite et ne sera jamais nulle même si les variables sont en réalité indépendantes. Comme l'information mutuelle estimée n'est jamais nulle, la valeur obtenue dépend fortement des données. Pour un autre ensemble de données de densité p^*, l'estimation de cette information mutuelle pourrait avoir une valeur différente et mener à un autre choix d'arêtes. Lorsque la dépendance entre les variable est forte, celle-ci s'exprime dans tout ensemble de données, contrairement à une faible dépendance. Pour réduire cette variance, certaines arêtes du graphe sont éliminées, l'arbre est donc ramené à un arbre non connecté avec k arêtes. Le nombre k d'arêtes est un paramètre de réglage qui induit un compromis entre le biais et la variance.

L'étape k de l'algorithme de Kruskal définit un ensemble $E^{(k)}$ d'arêtes qui correspond à la forêt $\hat{F}_{n_1}^{(k)}$ avec k arêtes ; où $\hat{F}_{n_1}^{(0)}$ est l'union de d sommets non connectés. Pour choisir k, l'article [18] suggère d'exploiter une validation croisée. Celle-ci est effectuée parmi les d forêts $\hat{F}_{n_1}^{(0)}, \hat{F}_{n_1}^{(1)}, \cdots, \hat{F}_{n_1}^{(d-1)}$ obtenues à chaque itération de l'algorithme de Kruskal. Ces forêts sont donc obtenues lors de la première étape qui consistait à construire une séquence de forêts à partir des observations du groupe $\mathcal{D}_1$.

Soient $\hat{p}_{ij,n_2}(x_i, x_j)$ et $\hat{p}_{k,n_2}(x_k)$ définis comme en (4.16) et en (4.17), mais maintenant évalués uniquement sur base des données conservées en $\mathcal{D}_2$. Le nombre d'arêtes conservées dans la forêt sera choisi afin de minimiser la distance entre la densité basée sur les observations de $\mathcal{D}_2$ et la densité basée sur les observations de $\mathcal{D}_1$ et supportée par la forêt $F_{n_1}^{(k)}$.

Comme remarqué précédemment, pour minimiser la distance entre deux fonctions de densité, il suffit de minimiser le risque négatif de la log-vraisemblance. Pour une densité p_F qui est supportée par une forêt F, le risque négatif conservé de la log-vraisemblance est défini comme

$$\begin{aligned}\hat{R}_{n_2}(p_F) &= -\int_{\chi} \hat{p}_{n_2}(x) \ln p_F(x) dx \\ &= -\sum_{(i,j)\in E_F} \int_{\chi_i \times \chi_j} \hat{p}_{ij,n_2}(x_i, x_j) \ln \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)} dx_i dx_j - \sum_{f \in V_F} \int_{\chi_f} \hat{p}_{f,n_2}(x_f) \ln p_f(x_f) dx_f.\end{aligned}$$

La forêt sélectionnée est alors celle minimisant le risque négatif conservé de la log-vraisemblance pour la densité $\hat{p}_{F_{n_1}^{(k)}}$ supportée par la forêt $F_{n_1}^{(k)}$. C'est donc la forêt $\hat{F}_{n_1}^{(\hat{k})}$ où

$$\hat{k} = \underset{k \in \{0, \cdots, d-1\}}{\arg\min} \hat{R}_{n_2}(\hat{p}_{F_{n_1}^{(k)}}) \tag{4.18}$$

et où $\hat{p}_{F_{n_1}^{(k)}}$ est calculé en utilisant l'estimation de la densité $\hat{p}_{n_1}$ construite sur $\mathcal{D}_1$.

Le deuxième membre de $\hat{R}_{n_2}(\hat{p}_{F_{n_1}^{(k)}})$ ne dépend pas des arêtes du graphe donc

$$\hat{k} = \underset{k\in\{0,\cdots,d-1\}}{\arg\min}\ \hat{R}_{n_2}(\hat{p}_{F_{n_1}^{(k)}}) \tag{4.19}$$

$$= \underset{k\in\{0,\cdots,d-1\}}{\arg\max} \sum_{(i,j)\in E_{F_{n_1}^{(k)}}} \int_{\chi_i\times\chi_j} \hat{p}_{ij,n_2}(x_i,x_j) \ln \frac{p_{ij,n_1}(x_i,x_j)}{p_{i,n_1}(x_i)p_{j,n_1}(x_j)} dx_i dx_j. \tag{4.20}$$

Une fois que $\hat{k}$ est obtenu, l'estimation finale de la densité estimée par noyau relative à la forêt est donnée par

$$\hat{p}_n(x) = \prod_{(i,j)\in E^{(\hat{k})}} \frac{\hat{p}_{ij,n_1}(x_i,x_j)}{\hat{p}_{i,n_1}(x_i)\hat{p}_{j,n_1}(x_j)} \prod_k \hat{p}_{k,n_1}(x_k). \tag{4.21}$$

4.4 En pratique

Pour appliquer cette technique, le logiciel R peut être utilisé. L'algorithme de Kruskal est déjà implémenté dans R via la fonction *msTreeKruskal* de la libraire *optrees*. Le reste de la procédure peut être implémenté manuellement, étape par étape. Voici la description de cette procédure sur R.

Pour commencer, les données doivent être séparées aléatoirement en deux groupes $\mathcal{D}_1$ et $\mathcal{D}_2$ de taille n_1 et n_2 respectivement. Après avoir effectué quelques recherches dans la littérature, aucune information concernant les tailles adéquates de chaque groupe n'a été trouvée. Cependant, dans l'article [22], lors d'un exemple numérique, le même nombre de données sont générées dans chaque groupe. C'est pourquoi, dans ce mémoire, les données seront séparées en deux groupes de tailles égales, c'est-à-dire $n_1 = n_2 = \frac{n}{2}$.

Pour estimer les densités univariées et bivariées à l'aide de la technique par noyau, à partir des données de $\mathcal{D}_1$, les formules (4.16) et (4.17) sont utilisées. Il existe différents types de noyau, mais, comme le choix du noyau influence peu le résultat de l'estimation, le noyau Gaussien est choisi. Pour le choix des tailles de fenêtre, la taille minimisant l'erreur quadratique moyenne intégrée sous des hypothèses gaussiennes est choisie. C'est-à-dire, pour l'estimation bivariée

$$h_{2k} = 1.06 \min\{\hat{\sigma}_k, \frac{\hat{a}_{k,0.75} - \hat{a}_{k,0.25}}{1.34}\} n_1^{-1/(2\beta+2)},$$

et

$$h_{1k} = 1.06 \min\{\hat{\sigma}_k, \frac{\hat{a}_{k,0.75} - \hat{a}_{k,0.25}}{1.34}\} n_1^{-1/(2\beta+1)},$$

pour l'estimation univariée où $\hat{\sigma}_{1k}$ est l'écart type empirique de $\{X_{sk}\}_{s\in\mathcal{D}_1}$ et $\hat{a}_{k,0.75}$ et $\hat{a}_{k,0.25}$ sont les quantiles empiriques 75% et 25%, avec $\beta = 2$. Des détails sur les noyaux et le choix de la taille de fenêtre sont repris dans le premier chapitre de [27].

Lorsque les densités marginales univariées et bivariées sont estimées, la matrice d'information mutuelle peut être estimée à son tour, voir (4.15). Il faut donc estimer une intégrale. Pour ce faire, l'estimation de la densité par noyau est calculée sur une grille de points. Il suffit de choisir m points d'évaluation dans chaque dimension χ_i, $x_{1i} < x_{2i} < \cdots < x_{mi}$ pour la variable X_i. L'information mutuelle $\hat{I}_{n_1}(X_i; X_j)$ est approximée par

$$\hat{I}_{n_1}(X_i; X_j) = \frac{1}{m^2} \sum_{k=1}^{m} \sum_{l=1}^{m} \hat{p}_{ij,n_1}(x_{ki}, x_{lj}) \ln \frac{\hat{p}_{ij,n_1}(x_{ki}, x_{lj})}{\hat{p}_{i,n_1}(x_{ki}) \hat{p}_{j,n_1}(x_{lj})}. \tag{4.22}$$

L'erreur d'approximation peut être rendue aussi petite que possible en choisissant m suffisamment grand. Le calcul de l'information mutuelle se fait en temps exponentiel par rapport à la valeur de m. Un bon compromis entre la précision de l'estimation et le temps de calcul est de prendre $m = 60$, lorsque le nombre de variables est compris entre 10 et 20.

Une fois que cette matrice d'information mutuelle est estimée, l'algorithme de Kruskal est appliqué. La fonction *msTreeKruskal* cherche l'arbre couvrant de poids minimal, cette fonction est donc appliquée à l'opposé de la matrice d'information mutuelle. Une fois l'algorithme appliqué, cette fonction renvoie une matrice dont les deux premières colonnes sont les sommets des arcs à inclure et dont la troisième colonne est le poids associé à cette arête. De plus, les arêtes sont ordonnées de sorte que la première ligne de la matrice contienne l'information de l'arête de poids minimal. Cela signifie qu'à l'itération i de l'algorithme est ajoutée l'arête de la ligne i.

Le graphe contenant l'arbre de poids maximum peut ensuite être représenté.

Il reste à effectuer l'étape 2 qui consiste à éliminer certaines arêtes si nécessaire. Pour ce faire, la formule (4.20) est utilisée. Pour obtenir, pour chaque $k \in \{0, \cdots, d\}$, les valeurs du risque, il faut estimer la valeur de l'intégrale. Comme précédemment, celle-ci sera estimée à l'aide d'une grille de points. Soient, dans chaque dimension χ_j, m points $x_{1j} < x_{2j} < \cdots < x_{mj}$, alors

$$\hat{k} = \underset{k \in \{0, \cdots, d-1\}}{\arg\max} \sum_{(i,j) \in E_{F_{n_1}^{(k)}}} \frac{1}{m^2} \sum_{l=1}^{m} \sum_{f=1}^{m} \hat{p}_{ij,n_2}(x_{li}, x_{fj}) \ln \frac{p_{ij,n_1}(x_{li}, x_{fj})}{p_{i,n_1}(x_{li}) p_{j,n_1}(x_{fj})}. \tag{4.23}$$

Il suffit ensuite de sélectionner les $\hat{k}$ premières arêtes reprises dans la matrice obtenue à la sortie de la fonction *msTreeKruskal*.

La forêt la plus représentative des données peut donc être représentée.

4.5 Exemples de modèles graphiques sous forme d'arbre

4.5.1 Exemple numérique

Avant d'effectuer le graphe créé à partir de la base de données, les exemples numériques du chapitre 2 sont repris. Ainsi, $n = 1000$ données sont générées selon la loi normale multivariée de dimension $d = 10$, de vecteur moyen le vecteur nul et dont la matrice de covariances a une structure particulière.

Comme, pour cette technique d'estimation du graphe, la matrice de concentration n'est plus utilisée, mais que l'estimation se base sur la log-vraisemblance, la standardisation des données n'est plus nécessaire.

Cas normal indépendant

Pour commencer, des données dont les variables sont indépendantes sont générées. Dans ce cas, il est attendu qu'aucune arête ne se trouve dans le graphe. C'est en effet le cas au vu de la Figure 4.3, qui représente les forêts obtenues avant (à gauche) et après (à droite) le choix optimal du nombre d'arêtes.

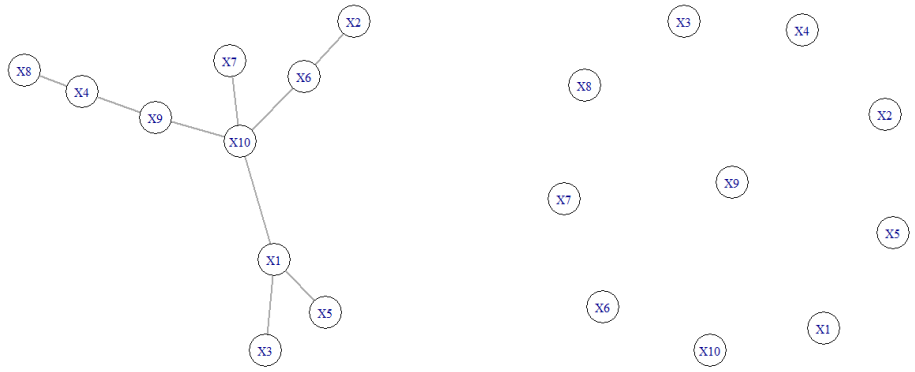

FIGURE 4.3 – Modèles graphique sous forme d'arbre pour des données générées selon une loi normale multivarivée de matrice de covariances la matrice identité. Le graphique de gauche représente l'arbre complet obtenu après la première étape d'estimation. Le graphe de droite est le graphe obtenu après la deuxième étape de l'estimation.

Cas normal équi-corrélé

Dans ce cas-ci, comme les variables ont toutes la même corrélation, l'information mutuelle entre chaque couple de variables devrait être presque identique. Il est difficile de savoir quelles arêtes vont être incluses dans le graphe. Il est néanmoins attendu que, soit toutes les arêtes de l'arbre soient conservées, soit qu'elles soient toutes supprimées, après l'étape 2 de la construction. Lorsque la corrélation entre les variables est de $\rho = 0.4$, les 9 arêtes de l'arbre sont conservées, voir le graphe de gauche de la Figure 4.4. La position des arêtes dépend, quant à elle, des données générées. Lorsque la corrélation entre les variables se rapproche de 0, le nombre d'arêtes conservées dans l'arbre diminue pour arriver à un graphe sans aucune arête ; c'est le cas lorsque $\rho = 0.1$. Lorsque $\rho = 0.2$, comme le montre le graphe de droite de la Figure 4.4, seules 6 arêtes sont conservées après l'étape 2 de l'estimation du graphe.

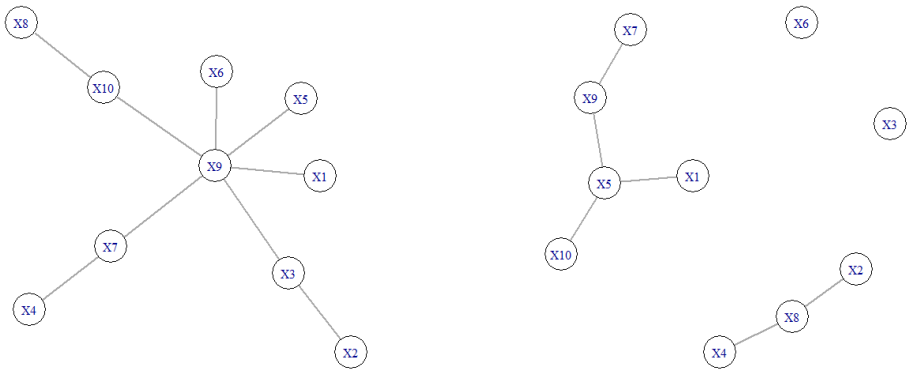

FIGURE 4.4 – Modèles graphiques sous forme d'arbre pour des données générées selon une loi normale multivarivée dont les variables sont équi-corrélées. Le graphe de gauche représente l'arbre obtenu après la deuxième étape d'estimation, lorsque la corrélation entre les variables est de $\rho = 0.4$. Le graphe de droite est l'arbre obtenu après la deuxième étape d'estimation, lorsque la corrélation entre les variables est de $\rho = 0.2$.

Cas normal auto-corrélé

L'étape suivante consiste à supposer que les variables sont liées par un taux de corrélation dépendant de la place des variables. Dans ce cas, comme le montre la Figure 4.5, reprenant l'arbre créé à l'aide de la technique présentée dans ce chapitre, les seules variables reliées par une arête sont les variables X_i et X_j telles que $|i-j|=1$. En effet, plus la distance entre i et j sera grande, moins la corrélation (donc la dépendance) entre les variables X_i et X_j sera élevée.

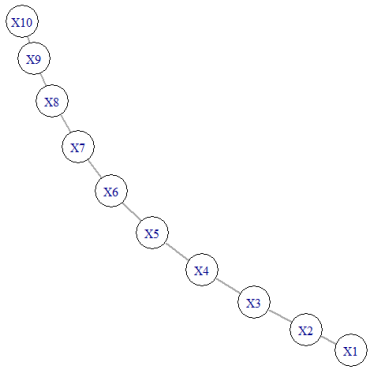

FIGURE 4.5 – Modèle graphique sous forme d'arbre pour des données auto-corrélées générées selon une loi normale avec $\rho = 0.4$. Le graphe représente l'arbre obtenu après la deuxième étape d'estimation, qui est le même que celui obtenu après la première étape.

Cas normal par bloc

Le dernier cas prend en compte des variables dont la matrice de covariances est une matrice par bloc. Comme précédemment, 10 variables sont considérées dont seules les 3 premières sont corrélées. Il est donc attendu que seules les arêtes reliant les variables X_1, X_2 et X_3 fassent partie du graphe. Or, comme le graphe ne peut pas comporter de cycle, une des 3 arêtes du triangle doit être supprimée. Ainsi, le graphe ne possèdera que deux arêtes, comme l'illustre la Figure 4.6, représentant les modèles graphiques pour cette structure de corrélation, obtenus après les première et deuxième étapes.

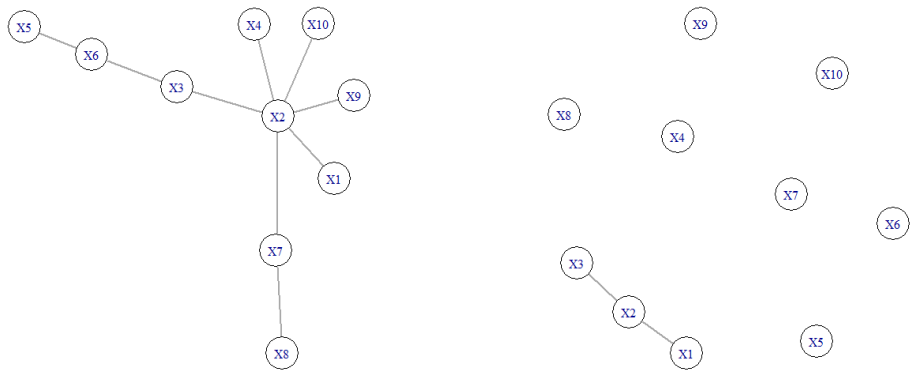

FIGURE 4.6 – Modèles graphiques sous forme d'arbre pour des données générées selon une loi normale multivarivée de matrice de covariances par bloc avec $\rho = 0.4$. Le graphique de gauche représente l'arbre complet obtenu après la première étape d'estimation. Le graphe de droite est le graphe obtenu après la deuxième étape de l'estimation.

Cas normal asymétrique

Pour considérer un cas non normal, la loi normale asymétrique est utilisée. Dans le chapitre 2 a été constaté que le graphe estimé à l'aide de la technique *glasso* possédait pas mal d'erreurs. Dans ce cas-ci, comme la technique ne demande aucune hypothèse particulière, si ce n'est que le graphe construit possède peu d'arêtes, le graphe estimé devrait être similaire au graphe attendu.

Pour effectuer cet exemple, 1000 données sont générées selon la loi normale asymétrique de paramètres

$$\Omega^{-1} = \begin{pmatrix} 1 & 0.4 & 0 & 0 & 0 \\ 0.4 & 1 & 0.4 & 0 & 0 \\ 0 & 0.4 & 1 & 0.4 & 0 \\ 0 & 0 & 0.4 & 1 & 0.4 \\ 0 & 0 & 0 & 0.4 & 1 \end{pmatrix}$$

et $\alpha = (0, 1, 0, 1, 0)$.

Comme vu dans le chapitre 2, les variables X_i et X_j sont indépendantes si et seulement si $\Omega^{-1}_{ij} = 0$ et $\alpha_i \alpha_j = 0$. C'est pourquoi le graphe représentant ces données devrait être le graphe représenté en Figure 4.7.

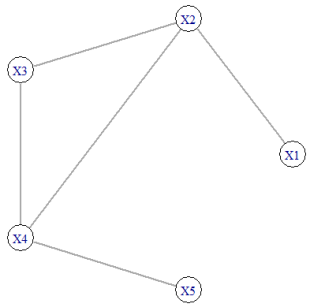

FIGURE 4.7 – Modèle graphique représentant théoriquement les données générées à partir de la loi normale asymétrique.

La matrice d'information mutuelle estimée sur ces données est la matrice

$$\begin{pmatrix} 0.0000 & 0.0013 & 0.0003 & 0.0004 & 0.0001 \\ 0.0013 & 0.0000 & 0.0003 & 0.0018 & 0.0003 \\ 0.0003 & 0.0003 & 0.0000 & 0.0005 & 0.0002 \\ 0.0004 & 0.0018 & 0.0005 & 0.0000 & 0.0016 \\ 0.0001 & 0.0003 & 0.0002 & 0.0016 & 0.0000 \end{pmatrix}.$$

Le graphe obtenu sur base de cette matrice est représenté en Figure 4.8. Celui-ci possède les mêmes arêtes que sur le graphe de la figure précédente comme attendu. Seule l'arête entre les variables X_2 et X_3 n'est pas reprise car celle-ci causerait un cycle et le graphe construit ne serait plus un arbre.

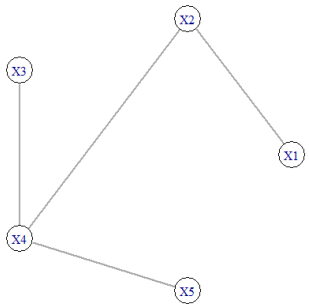

FIGURE 4.8 – Modèle graphique, tel qu'obtenu à l'aide de la technique d'estimation sous forme d'arbre, représentant les données générées à partir de la loi normale asymétrique.

Cas de dépendances non linéaires

Pour cet exemple, les variables créées pour l'exemple contenant des dépendances non linéaires du chapitre 2 sont reprises. Pour rappel, les différentes variables ont été générées à l'aide des définitions suivantes :

- $X_1 \leftarrow \exp(1)$
- $X_2 \leftarrow X_1^2$
- $X_3 \leftarrow \sqrt{X_1}$
- $X_4 \leftarrow$ unif $(-2, 2)$
- $X_5 \leftarrow X_4^2$
- $X_6 \leftarrow X_4^3$
- $X_7 \leftarrow X_3 + X_5$
- $X_8 \leftarrow X_4 X_1$

Pour commencer l'estimation de la structure du graphe, il faut tout d'abord estimer les densités univariées et bivariées des variables. En Figure 4.9 se trouvent les estimations des fonctions de densité univariées pour les 8 variables. En Figure 4.10 se trouvent les estimations des fonctions de densité bivariées pour certains couples de variables.

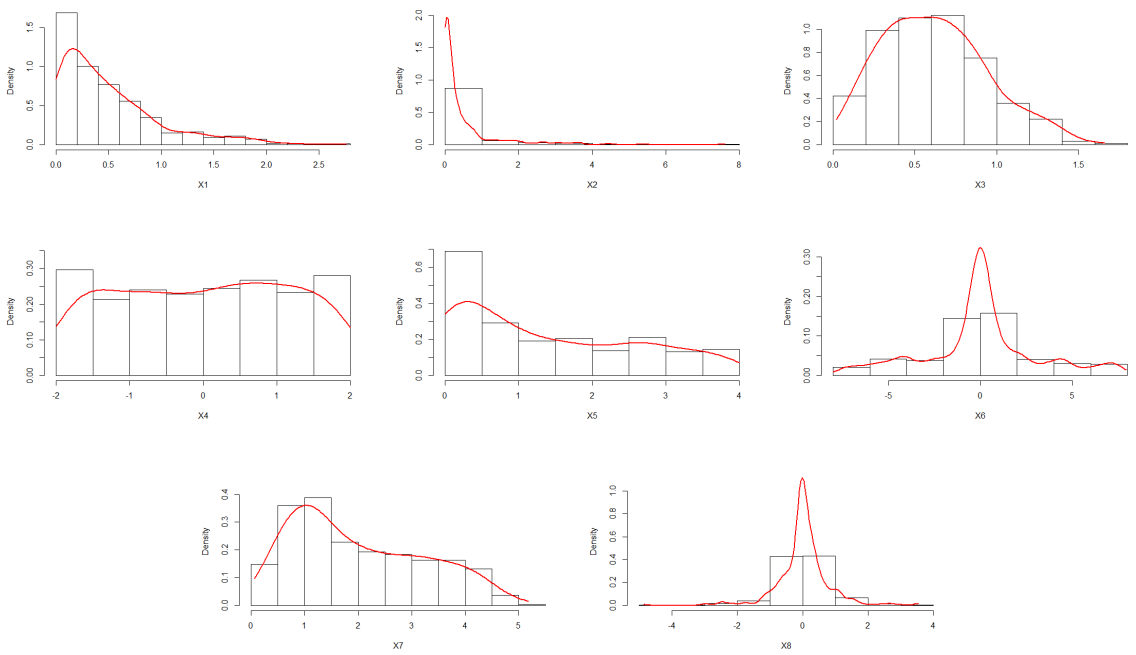

FIGURE 4.9 – Histogrammes des différentes variables sur lesquels l'estimation de la densité univariée de chaque variable est représentée en rouge.

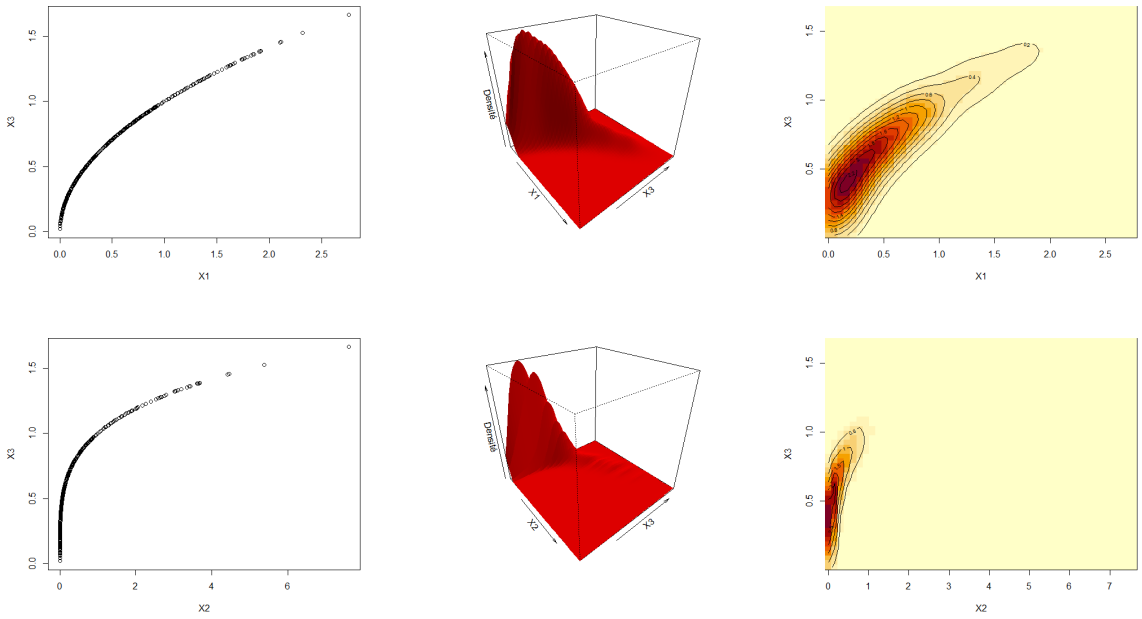

FIGURE 4.10 – Estimation de fonctions de densité bivariées.

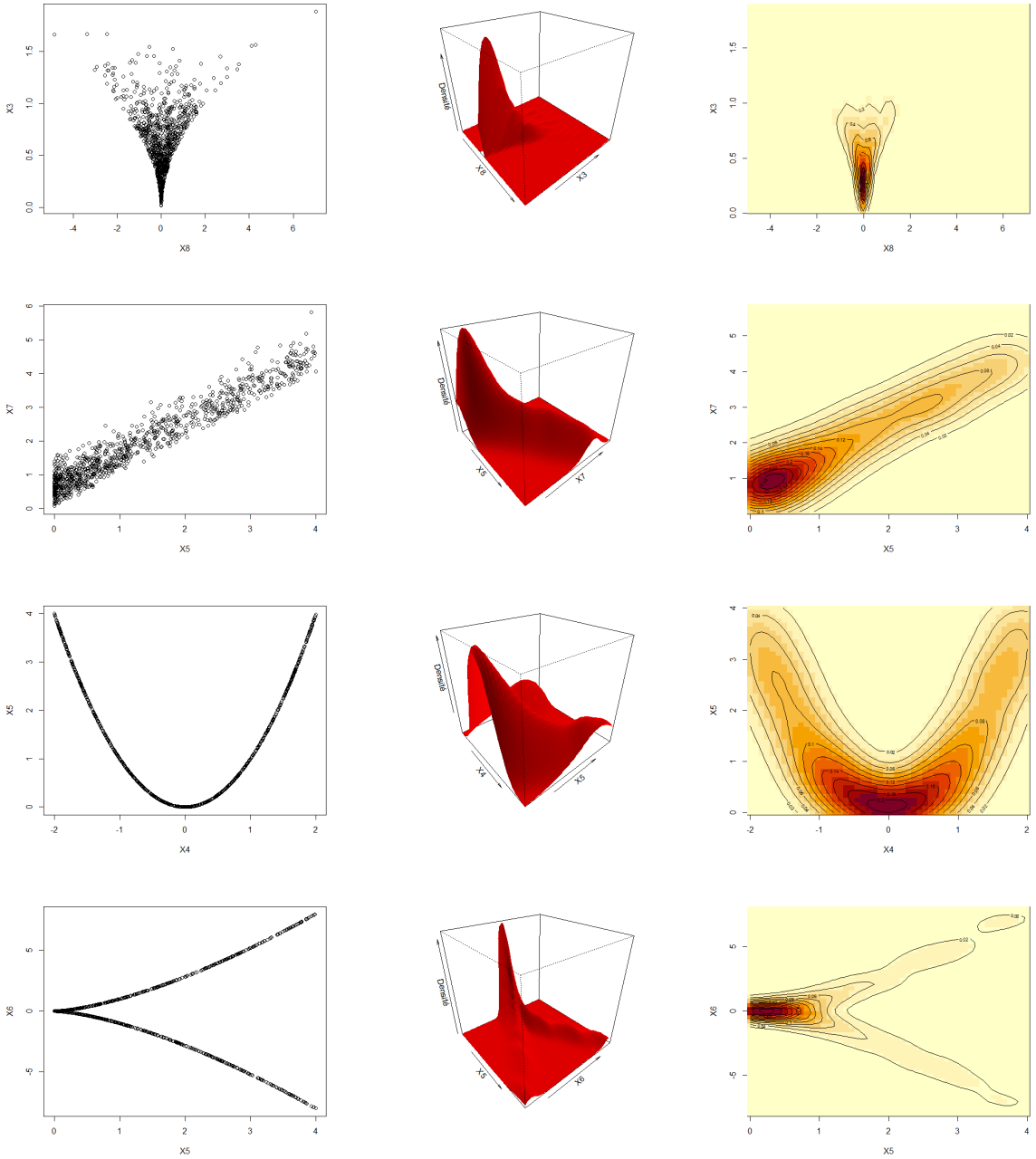

FIGURE 4.10 – Estimation de fonctions de densité bivariées.

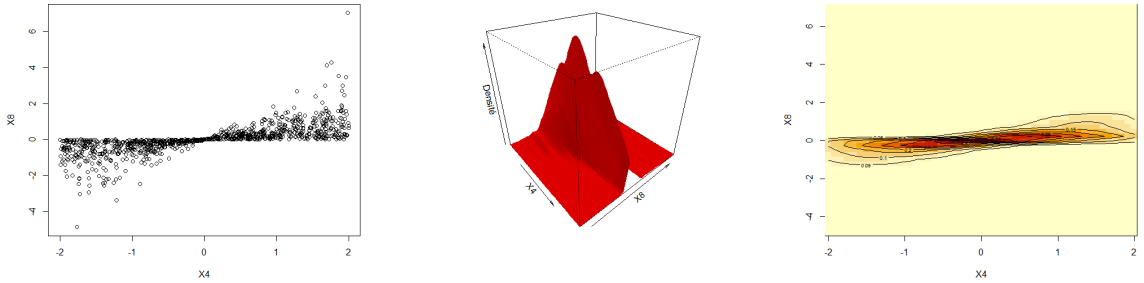

FIGURE 4.10 – Estimation de fonctions de densité bivariées.

La matrice d'information mutuelle est alors estimée et celle-ci est égale à la matrice

$$\begin{pmatrix} 0 & 0.0287 & 0.1222 & -0.0020 & -0.0022 & 0.0001 & 0.0003 & 0.0120 \\ 0.0287 & 0 & 0.0418 & -0.0006 & -0.0006 & 0.0002 & 0.0002 & 0.0047 \\ 0.1222 & 0.0418 & 0 & -0.0024 & -0.0026 & 0.0005 & 0.0014 & 0.0227 \\ -0.0020 & -0.0006 & -0.0024 & 0 & 0.0114 & 0.0087 & 0.0104 & 0.0104 \\ -0.0022 & -0.0006 & -0.0026 & 0.0114 & 0 & 0.0090 & 0.0212 & 0.0018 \\ 0.0001 & 0.0002 & 0.0005 & 0.0087 & 0.0090 & 0 & 0.0070 & 0.0028 \\ 0.0003 & 0.0002 & 0.0014 & 0.0104 & 0.0212 & 0.0070 & 0 & 0.0041 \\ 0.0120 & 0.0047 & 0.0227 & 0.0104 & 0.0018 & 0.0028 & 0.0041 & 0 \end{pmatrix}.$$

Les valeurs négatives de la matrice peuvent être vues comme des valeurs nulles. En effet, l'information mutuelle entre deux variables est toujours positive et sera nulle si ces variables sont indépendantes. Comme les fonctions de densité univariées et bivariées sont estimées, tout comme l'information mutuelle, il se peut qu'une valeur proche d'une valeur nulle soit estimée par une valeur négative. Les variables X_1, X_2 et X_3 sont bien indépendantes des variables X_4 et X_5, au vu de la façon dont elles ont été créées.

L'arbre obtenu sur base de cette matrice possède alors 7 arêtes et est représenté en Figure 4.11. De plus, aucune arête ne doit être enlevée après la deuxième étape de l'estimation. Ainsi, le graphe final est bien le graphe de la Figure 4.11.

Les couples de variables repris dans les estimations des fonctions de densité bivariées représentées en Figure 4.10 sont en fait choisis comme étant les couples de variables reliées dans l'arbre représentant les données. Il est clair, au vu des graphiques de la première colonne de la Figure 4.10, que ces couples de variables sont dépendants.

Pour comparer l'arbre obtenu dans ce chapitre avec le graphe obtenu sur les mêmes données dans le chapitre 2, seules les arêtes connectant les variables X_1 et X_3, X_2 et X_3, X_5 et X_7 ainsi que X_4 et X_8 sont communes aux deux graphes. Cela est du au fait qu'il est possible d'envisager une dépendance linéaire entre ces couples de variables. Les trois autres couples de variables qui sont reliés dans l'arbre possèdent une dépendance qui est autre que linéaire.

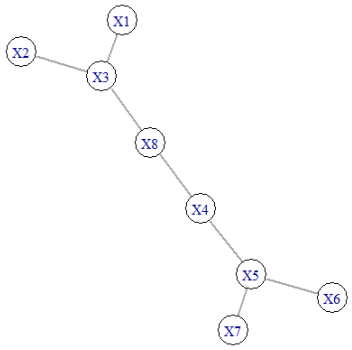

FIGURE 4.11 – Arbre estimé sur base des données non linéaires.

4.5.2 Modèle graphique sous forme de forêt des données socio-économiques de la Wallonie

Il reste maintenant à construire le graphe caractérisant les données socio-économiques de la Wallonie. Avant tout, la première étape consiste à estimer les fonctions de densité univariées et bivariées. En Figure 4.12 se trouvent 4 exemples de l'estimation des densités marginales univariées. Les histogrammes des variables sont représentés ainsi que, en rouge, la fonction de densité estimée par noyau. Les fonctions estimées respectent bien la forme des histogrammes. Deux exemples d'estimation des densités bivariées sont aussi repris en Figure 4.13.

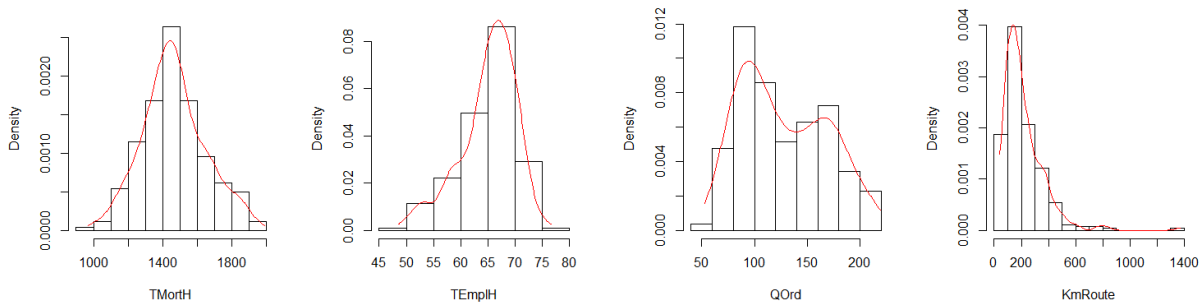

FIGURE 4.12 – Exemples d'estimation de densités univariées par noyau. L'histogramme est créé à partir des données du groupe $\mathcal{D}_1$ et la ligne rouge est la fonction de densité estimée sur ces données.

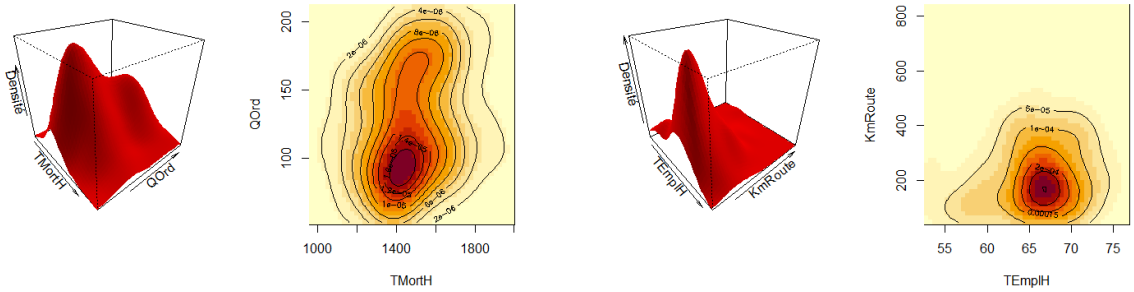

FIGURE 4.13 – Exemples d'estimation de densités bivariées par noyau. Ces estimations sont basées sur les données du groupe $\mathcal{D}_1$.

La prochaine étape est d'estimer la matrice d'information mutuelle. Une fois qu'elle est estimée, les modèles graphiques peuvent à leur tour être estimés. Le graphe obtenu après la première étape possèdera 18 arêtes, alors que celui obtenu après la deuxième étape en possèdera moins.

A l'aide de la Figure 4.14, représentant les modèles graphiques pour la base de données obtenus grâce à la technique présentée dans ce chapitre, il est possible de voir, sur le graphe de droite, les variables ayant les dépendances les plus fortes. Ainsi, comme déjà remarqué, à priori, les variables concernant les taux d'emploi et de chômage pour les hommes et pour les femmes sont fortement dépendantes. De plus celles-ci sont aussi fortement liées aux variables TPartElec, AgeMoy et TNat. Ces dernières variables ne semblaient pas, à priori, faire partie des dépendances les plus élevées. Il est intéressant de remarquer aussi que les arêtes liant ces trois dernières variables ne faisaient pas partie des graphes possédant 18 arêtes obtenus à partir des techniques *non paranormale* et *glasso*.

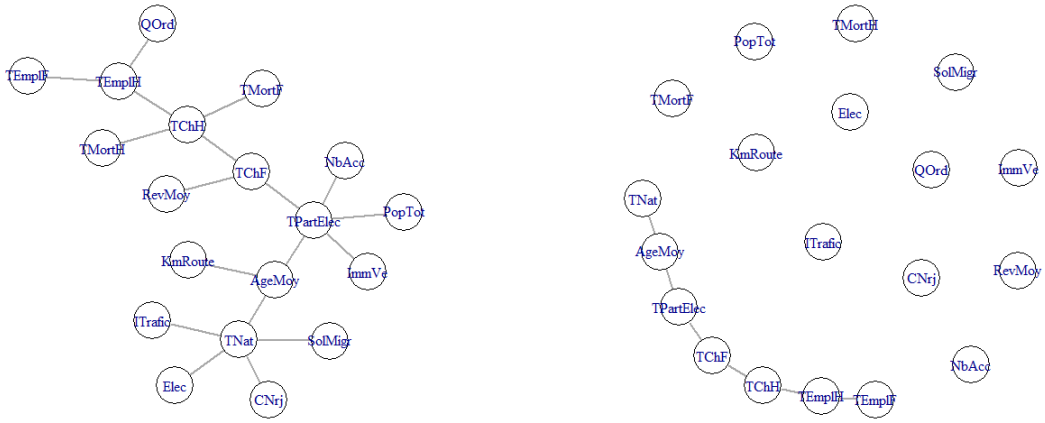

FIGURE 4.14 – Modèles graphiques sous forme d'arbre pour la base de données. Le graphique de gauche représente l'arbre complet obtenu après la première étape d'estimation. Le graphe de droite est le graphe obtenu après la deuxième étape de l'estimation.

Chapitre 5

Comparaison des techniques présentées

Après avoir présenté trois différentes techniques d'estimation de modèles graphiques, il est intéressant de comparer leur efficacité. Cette efficacité peut être mesurée, notamment, par le taux d'erreur. L'étude théorique de l'erreur commise dans la construction des modèles graphiques fait l'objet du cours de "Research problems in probability and statistics". Dans ce mémoire, seul l'aspect pratique sera étudié. Pour estimer le taux d'erreur des différentes techniques d'estimation de structure de graphes, seule la présence ou l'absence des arêtes dans le graphe seront étudiées ; ce qui permet de calculer les taux de faux positif, de faux négatif et d'erreur. Il pourrait aussi être intéressant de comparer, à leur vraie valeur, les estimations des valeurs reprises dans la matrice Θ pour les techniques des deux premiers chapitres et dans la matrice d'information mutuelle pour la dernière technique, mais cela ne fera pas l'objet de ce mémoire.

Les trois techniques comparées sont les suivantes :

— Technique 1 : *glasso* ;

— Technique 2 : *non paranormal* (NP) ;

— Technique 3 : *arbre*.

Les types d'erreur "faux positif" et "faux négatif" sont respectivement définis par la conservation d'une arête dans le graphe à tort et la suppression d'une arête à tort. Pour mesurer les types d'erreur les plus présents dans l'estimation de la structure d'un graphe, il est classique de calculer les taux de faux positif et de faux négatif tout comme le taux d'erreur. Ces deux types d'erreur seront considérées de façon symétrique car ils ont la même importance.

Soient $X = (X_1, \cdots, X_d)$ un vecteur aléatoire et $\mathcal{G}$ le graphe estimé sur base de ce vecteur. Soit Y la matrice $d \times d$ définie par

$$Y_{ij} = \begin{cases} 0 & \text{si } X_i \amalg X_j | X_{\{i,j\}} \\ 1 & \text{sinon.} \end{cases}$$

Avoir une erreur de type faux positif (FP) signifie que la matrice d'adjacence A du graphe $\mathcal{G}$ possède un 1 ($A_{ij} = 1$) alors que les variables X_i et X_j sont indépendantes ($Y_{ij} = 0$). Dans le cas de l'erreur de type faux négatif (FN), $Y_{ij} = 1$ et $A_{ij} = 0$. Après avoir effectué quelques recherches dans la littérature ([28], [3]), il semble judicieux de définir le taux d'erreur comme étant le nombre de couples de variables (X_i, X_j), $1 \leq i < j \leq d$ pour lesquelles la valeur de Y_{ij} est différente de la valeur de A_{ij} et ce, divisé par le nombre total de couples ; c'est-à-dire

$$\begin{aligned}
\text{taux d'erreur } &= \frac{2}{d(d-1)} \sum_{1 \leq i < j \leq d} I(Y_{ij} \neq A_{ij}) \\
&= \frac{2}{d(d-1)} \left(\sum_{1 \leq i < j \leq d} I(Y_{ij} = 0 \text{ et } A_{ij} = 1) + \sum_{1 \leq i < j \leq d} I(Y_{ij} = 1 \text{ et } A_{ij} = 0) \right) \\
&= \frac{2}{d(d-1)} (\text{ nombre de FP } + \text{ nombre de FN }).
\end{aligned}$$

Il suffit de considérer uniquement le triangle inférieur des deux matrices puisque celles-ci sont symétriques. De plus, la diagonale ne doit pas être prise en compte car le graphe ne possèdera jamais de boucle sur un même sommet.

5.1 Taux d'erreur sous la loi normale

Pour commencer cette étude d'erreur, les données sont supposées suivre une loi normale. Dans le cas de cette loi, les trois techniques devraient estimer correctement la structure du graphe, même si quelques erreurs sont tout de même possibles. En effet, l'étude des différents exemples numériques du chapitre 2 a permis de constater que lorsque les critères AIC et BIC sont utilisés pour choisir le paramètre de pénalisation de la technique *glasso*, certaines erreurs sont commises. Ce problème est aussi présent pour la technique *non paranormale* car les mêmes critères sont utilisés. Pour la troisième technique, les arêtes placées dans le graphe devraient relier des variables qui sont effectivement dépendantes. Cependant, comme le nombre d'arêtes présentes dans le graphe est très restreint (maximum $d-1$ arêtes), le nombre de faux négatifs pourrait grimper lorsque le nombre de couples de variables dépendantes est élevé.

Pour estimer le taux d'erreur des différentes techniques, $n = 200$ données sont générées selon la loi $\mathcal{N}_d(\mu, \Sigma)$ avec $\Theta = \Sigma^{-1}$. Pour rappel, la matrice Y peut, dans ce cas, être définie par

$$Y_{ij} = \begin{cases} 0 & \text{si } \Theta_{ij} = 0 \\ 1 & \text{sinon.} \end{cases}$$

Les taux de faux positif, de faux négatif et d'erreur peuvent être calculés sur base de cette matrice et de la matrice d'adjacence du graphe estimé.

Il est intéressant de considérer des exemples pour lesquels la matrice de concentration Θ a une structure connue. Pour chacun de ces exemples, à chaque itération, un ensemble

de $n = 200$ données seront générées selon une loi $\mathcal{N}_d(0, \Theta^{-1})$, où $d = 10$ et Θ varie en fonction des exemples. Pour chaque type d'exemple, $N = 100$ itérations seront effectuées et pour chacune d'elles, 3 éléments seront conservés : le taux de faux positif (FP), le taux de faux négatif (FN) et le taux d'erreur. Ensuite, pour chacun de ces éléments, la moyenne des 100 itérations sera calculée. Les différents exemples sont les suivants :

— Cas indépendant : $\Theta = I_{10}$.

— Cas équi-corrélé : $\Theta = \Sigma_{0.4,10}$.

— Cas par bloc 1 : $\Theta = \begin{pmatrix} \Sigma_{0.4,8} & 0_{8,2} \\ 0_{2,8} & I_2 \end{pmatrix}$.

— Cas par bloc 2 : $\Theta = \begin{pmatrix} \Sigma_{0.4,5} & 0_{5,5} \\ 0_{5,5} & I_5 \end{pmatrix}$.

— Cas par bloc 3 $\Theta = \begin{pmatrix} \Sigma_{0.4,9} & 0_{9,1} \\ 0_{1,9} & 1 \end{pmatrix}$.

La différence entre les 3 cas par bloc est le nombre de 0 dans le triangle inférieur de la matrice de concentration. Soit s le taux de valeurs nulles dans la matrice de concentration. Ce taux est calculé sur base du triangle inférieur de la matrice et sans tenir compte de la diagonale. Pour le premier cas, le triangle inférieur possède 17 zéros, c'est à dire 40% ($s = 0.4$) de valeurs nulles. Pour les deux et troisième cas, ceux-ci possèdent 35 et 9 zéros respectivement, c'est-à-dire environ 80% ($s = 0.8$) et 20% ($s = 0.2$) de valeurs nulles.

Pour la technique *glasso*, les données générées sont, bien entendu, standardisées comme imposé dans la partie théorique. Pour les simulations des techniques 1 et 2, la valeur du paramètre de pénalisation est choisie de sorte à minimiser le critère BIC. La Table 5.1 reprend les valeurs moyennes des taux de faux positif, de faux négatif et d'erreur pour les différents exemples.

Le fait que le nombre d'erreurs soit fortement dû au nombre de faux positifs, pour tous les exemples, pour les deux premières techniques, rejoint les conclusions tirées à partir des exemples numériques du chapitre 2. En effet, il en ressortait que le nombre d'arêtes conservées dans le graphe est supérieur au nombre de variables réellement dépendantes, donc que certaines arêtes du graphe étaient conservées à tort. C'est pourquoi aussi, dans le cas équi-corrélé, aucune erreur n'est commise car le graphe attendu est un graphe complet et aucune arête ne peut être ajoutée en plus. Dans le cas indépendant, très peu d'arêtes sont conservées à tort. Au vu du tableau, pour les techniques 2 et 3, le taux d'erreur est plus élevé lorsque le taux de variables indépendantes est de 0.4, que s'il est de 0, 0.2, 0.8 ou 1.

$d = 10$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0.0337	0	0.0337
		NP	0.0320	0	0.0320
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0	0	0
		NP	0	0	0
		arbre	1	1	0
Cas par bloc 1	0.4	*glasso*	0.3688	0.0275	0.3413
		NP	0.3811	0.0662	0.3148
		arbre	0.6220	0.6220	0
Cas par bloc 2	0.8	*glasso*	0.2344	0.0117	0.2226
		NP	0.2311	0.0133	0.2177
		arbre	0.2220	0.2220	0
Cas par bloc 3	0.2	*glasso*	0.1993	0	0.1993
		NP	0.2035	0.00080	0.1955
		arbre	0.8000	0.8000	0

TABLE 5.1 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 10$.

Pour la dernière technique, le taux de faux positif est toujours nul, ce qui signifie que les arêtes ajoutées au graphe relient des variables qui sont effectivement dépendantes. Cependant, comme le nombre d'arêtes qu'il est possible d'ajouter au graphe avec cette technique est très faible, le nombre de faux négatifs est très élevé. Cette technique possède surtout des avantages lorsque le taux de variables indépendantes s est très élevé. En effet, pour le cas par bloc 2, le taux d'erreur de la technique d'estimation sous forme d'arbre est plus faible que pour les deux autres techniques, tout comme dans le cas indépendant.

Il est aussi intéressant de comparer ces techniques pour des données de dimension plus grande. Pour ce faire, les mêmes simulations sont effectuées pour $d = 20$. Les différents exemples sont les suivants :

— Cas indépendant : $\Theta = I_{20}$.

— Cas équi-corrélé : $\Theta = \Sigma_{0.4,20}$.

— Cas par bloc 1 : $\Theta = \begin{pmatrix} \Sigma_{0.4,15} & 0_{15,5} \\ 0_{5,15} & \Sigma_{0.4,5} \end{pmatrix}$.

— Cas par bloc 2 : $\Theta = \begin{pmatrix} \Sigma_{0.4,9} & 0_{9,11} \\ 0_{11,9} & I_{11} \end{pmatrix}$.

— Cas par bloc 3 $\Theta = \begin{pmatrix} \Sigma_{0.2,18} & 0_{18,2} \\ 0_{2,18} & I_2 \end{pmatrix}$.

Dans cette dimension, le nombre de paramètres présents dans le triangle inférieur de la matrice de concentration est de 190. Le cas par bloc 1 (resp. 2 et 3) possède 75 (resp. 154 et 37) zéros dans la demi matrice de concentration, ce qui signifie que $s = 0.4$ (resp. $s = 0.8$ et $s = 0.2$). Les résultats obtenus sur base de ces différents exemples sont repris dans la Table 5.2.

$d = 20$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0.0155	0	0.0155
		NP	0.0151	0	0.0151
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.9807	0.9807	0
		NP	0.9810	0.9810	0
		arbre	0.9998	0.9998	0
Cas par bloc 1	0.4	*glasso*	0.5646	0.5613	0.0032
		NP	0.5685	0.5655	0.0030
		arbre	0.6051	0.6051	0
Cas par bloc 2	0.8	*glasso*	0.1753	0.1687	0.0065
		NP	0.1759	0.1686	0.0072
		arbre	0.1894	0.1894	0
Cas par bloc 3	0.2	*glasso*	0.7861	0.7845	0.0016
		NP	0.7854	0.7839	0.0014
		arbre	0.8052	0.8052	0

TABLE 5.2 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 20$.

Pour les deux premières techniques, les résultats sont fort différents de ceux obtenus en dimension 10. La principale cause de cette différence est que, dans cette dimension, la valeur du paramètre de pénalisation est fort élevée et le graphe estimé possède trop peu d'arêtes. Il semble donc qu'utiliser le critère BIC pour trouver une valeur adéquate du paramètre de pénalisation mène à une variation des types d'erreur selon la dimension. Dans cette dimension, le taux d'erreur augmente lorsque s diminue, alors qu'en dimension 10, celui-ci était maximal au niveau de $s = 0.4$.

Comme première conclusion, le modèle graphique obtenu à partir de la méthode *non paranormale* est moins utile dans le cas de données suivant une loi multinormale car son taux d'erreur est toujours supérieur à celui estimé pour la méthode *glasso*. De plus, il semble judicieux de choisir le paramètre de pénalisation des deux premières techniques d'une autre manière qu'à l'aide du critère BIC. Le critère AIC ne permettra pas d'obtenir des résultats plus représentatifs car il est basé sur le même type de calcul.

5.1.1 Changement du paramètre de pénalisation

Etant donné que choisir le paramètre de pénalisation en fonction des critères AIC et BIC ne semble pas adéquat, une autre façon de choisir ce paramètre pourrait peut-être faire diminuer le nombre d'erreurs commises pour les deux premières techniques. Dans l'article [21], la valeur du paramètre de pénalisation est choisie de sorte à minimiser le taux d'erreur. Ce choix sera aussi adopté dans ce mémoire pour les différents exemples qui suivent. Pour trouver le paramètre de pénalisation minimisant l'erreur, les taux d'erreur des graphes estimés pour λ allant de 0.01 à 1 par pas de 0.01 ont été calculés et la valeur de λ permettant de minimiser le taux d'erreur ainsi que le graphe estimé pour cette valeur sont conservés.

Les moyennes des taux d'erreur, de faux positif et de faux négatif sont uniquement recalculées, pour les deux premières techniques car la troisième reste inchangée. Les Table 5.3 et Table 5.4 contiennent ces valeurs pour les différents exemples de la section précédente, pour les dimensions $d = 10$ et $d = 20$.

$d = 10$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
		glasso	0	0	0
Cas indépendant	1	*NP*	0	0	0
		glasso	0.0106	0.0106	0
Cas équi-corrélé	0	*NP*	0.0133	0.0133	0
		glasso	0.1777	0.0402	0.1375
Cas par bloc 1	0.4	*NP*	0.1815	0.0400	0.1415
		glasso	0.0660	0.0340	0.0320
Cas par bloc 2	0.8	*NP*	0.0717	0.0395	0.0322
		glasso	0.1195	0.0104	0.1091
Cas par bloc 3	0.2	*NP*	0.1244	0.0097	0.1146

TABLE 5.3 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 10$.

Choisir le paramètre de cette façon permet de diminuer fortement le taux d'erreur de ces deux premières techniques. Cela montre qu'il est donc possible d'estimer un modèle graphique représentant une base de données dont les arêtes sont placées correctement avec un taux de confiance suffisamment élevé. De plus, le taux d'erreur a cette fois le même comportement en fonction de s, que $d = 10$ ou $d = 20$. Celui-ci est plus élevé lorsque $s = 0.4$. Le taux d'erreur augmente cependant lorsque la dimension augmente et que le nombre d'observations dans la base de données reste inchangé.

$d = 20$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
		glasso	0	0	0
Cas indépendant	1	*NP*	0	0	0
		glasso	0.0501	0.0501	0
Cas équi-corrélé	0	*NP*	0.0553	0.0553	0
		glasso	0.3047	0.0450	0.2597
Cas par bloc 1	0.4	*NP*	0.3069	0.0488	0.2581
		glasso	0.1593	0.1341	0.0251
Cas par bloc 2	0.8	*NP*	0.1606	0.1351	0.0255
		glasso	0.2025	0.0328	0.1696
Cas par bloc 3	0.2	*NP*	0.2086	0.0411	0.1675

TABLE 5.4 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 20$.

5.1.2 Taux d'erreur en fonction du taux de variables indépendantes

Il est intéressant ensuite d'étudier le taux d'erreur des trois méthodes lorsque la structure de Θ ainsi que les valeurs composant la matrice sont aléatoires. Cela permet de généraliser l'étude du taux d'erreur en fonction du taux de variables indépendantes. Pour ce faire, à chaque itération, une matrice de concentration Θ est construite de sorte qu'elle contienne un certain taux s de valeurs nulles. Ensuite des données sont générées selon la loi $\mathcal{N}_d(0, \Theta^{-1})$ et les trois techniques sont appliquées. La construction de la matrice de concentration est basée sur la construction effectuée dans [26]. Celle-ci est constituée des étapes suivantes :

1. Soit s le taux de valeurs nulles dans Θ : $s = \text{Unif}[0, 1]$.
2. Pour tout $i < j$,

$$\Theta_{ij} = \begin{cases} \text{Unif}]0.25, 0.75[& \text{avec une probabilité } (1-s) \times 0.75 \\ \text{Unif}]-0.75, -0.25[& \text{avec une probabilité } (1-s) \times 0.25 \\ 0 & \text{avec une probabilité } s \end{cases}$$

3. Pour tout $i < j$, $\Theta_{ij} = \Theta_{ji}$ pour obtenir la symétrie et $\Theta_{ii} = 1$ pour tout i.
4. Pour que Θ soit définie positive, poser, pour tout i,

$$\Theta_{ii} = 1 - e_{min} + 0.1 \text{ si } e_{min} \leq 0,$$

où e_{min} est la valeur propre minimale de Θ.

Pour effectuer ces simulations, un nombre $N = 200$ d'itérations sont effectuées pour lesquelles $n = 200$ données sont générées selon la loi $\mathcal{N}_{10}(0, \Theta^{-1})$. Lors de chaque itération,

le taux d'erreur, la valeur de s, le taux de faux positif et le taux de faux négatif sont calculés.

Voici deux exemples de matrice de concentration créés en dimension $d = 10$ avec la procédure ci-dessus. La première matrice possède un taux de valeurs nulles égal à 0.4 :

$$\begin{pmatrix} 1.70 & 0 & 0.58 & 0.32 & 0 & -0.26 & 0.36 & 0 & 0 & 0.54 \\ 0 & 1.70 & 0.58 & 0.68 & 0.67 & 0.45 & 0 & -0.58 & 0 & -0.41 \\ 0.58 & 0.58 & 1.70 & 0 & 0.65 & 0 & 0.54 & 0.40 & -0.74 & 0 \\ 0.32 & 0.68 & 0 & 1.70 & 0 & -0.29 & -0.62 & 0 & 0.66 & 0.43 \\ 0 & 0.67 & 0.65 & 0 & 1.70 & 0.59 & 0.64 & 0.53 & 0 & -0.25 \\ -0.26 & 0.45 & 0 & -0.29 & 0.59 & 1.70 & 0 & 0 & -0.66 & 0 \\ 0.36 & 0 & 0.54 & -0.62 & 0.64 & 0 & 1.70 & 0 & 0.31 & -0.75 \\ 0 & -0.58 & 0.40 & 0 & 0.53 & 0 & 0 & 1.70 & 0 & 0 \\ 0 & 0 & -0.74 & 0.66 & 0 & -0.66 & 0.31 & 0 & 1.70 & -0.41 \\ 0.54 & -0.41 & 0 & 0.43 & -0.25 & 0 & -0.75 & 0 & -0.41 & 1.70 \end{pmatrix}. \quad (5.1)$$

La deuxième présente un taux de valeurs nulles égal à 0.7 :

$$\begin{pmatrix} 1.84 & 0.29 & 0 & 0 & 0.32 & 0 & 0.59 & 0 & 0 & 0 \\ 0.29 & 1.84 & 0 & -0.58 & 0.42 & 0.47 & 0.31 & 0.73 & 0.67 & 0 \\ 0 & 0 & 1.84 & -0.56 & 0 & 0.73 & 0.67 & 0 & 0 & 0 \\ 0 & -0.58 & -0.56 & 1.84 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.32 & 0.42 & 0 & 0 & 1.84 & 0 & -0.32 & -0.70 & 0 & 0.28 \\ 0 & 0.47 & 0.73 & 0 & 0 & 1.84 & 0 & 0 & 0 & 0 \\ 0.59 & 0.31 & 0.67 & 0 & -0.32 & 0 & 1.84 & 0 & 0 & 0 \\ 0 & 0.73 & 0 & 0 & -0.70 & 0 & 0 & 1.84 & 0 & 0 \\ 0 & 0.67 & 0 & 0 & 0 & 0 & 0 & 0 & 1.84 & 0 \\ 0 & 0 & 0 & 0 & 0.28 & 0 & 0 & 0 & 0 & 1.84 \end{pmatrix}. \quad (5.2)$$

En Figure 5.1 se trouvent les boites à moustaches représentant les estimations des taux d'erreur, de faux positif et de faux négatif pour les trois techniques présentées. Sur base de ces graphiques, il est facile de constater que pour les deux premières techniques, les erreurs commises sont davantage du type faux positif, alors qu'elles sont plutôt du type faux négatif pour la technique d'estimation sous forme d'arbre. De plus, les taux d'erreur des techniques *glasso* et *non paranormale* sont presque identiques. La médiane du taux d'erreur pour la deuxième technique est tout de même légèrement supérieure à celle de la première technique. Ce qui signifie, à nouveau, que, lorsque les données suivent une loi multinormale, il n'est pas nécessaire d'effectuer les transformations de la technique *non paranormale*.

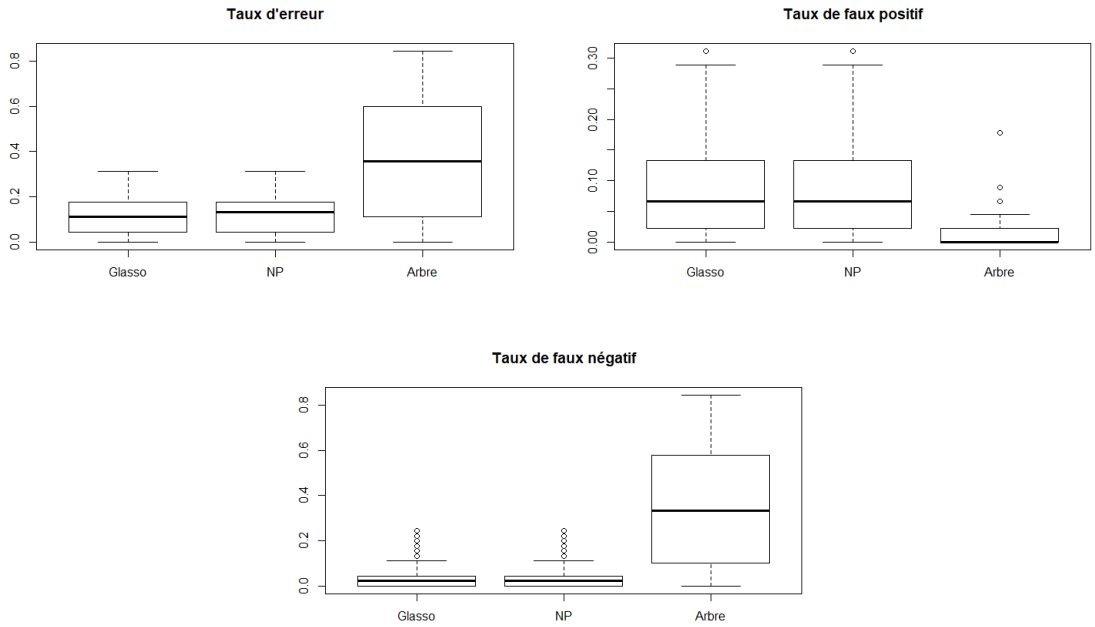

FIGURE 5.1 – Boites à moustaches représentant les estimations des taux d'erreur, de faux positif et de faux négatif dans les graphes obtenus à l'aide des trois techniques présentées. Les simulations sont effectuées pour $n = 200$ et $d = 10$.

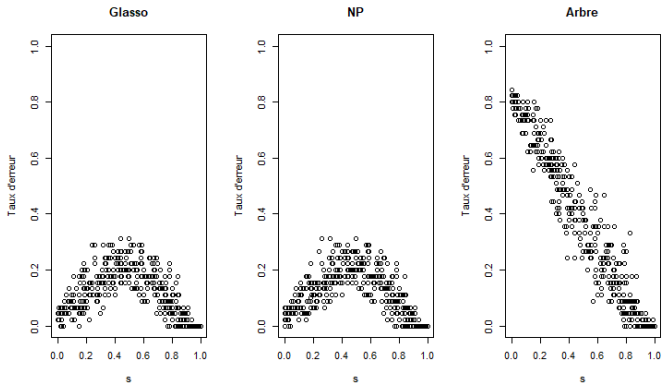

FIGURE 5.2 – Graphiques représentant le taux d'erreur en fonction du taux de variables indépendantes s pour des modèles graphiques estimés à l'aide des trois techniques présentées. Les simulations sont effectuées pour $n = 200$ et $d = 10$.

La Figure 5.2 représentant le taux d'erreur en fonction du taux de variables indépendantes s pour des modèles graphiques estimés à l'aide des trois techniques présentées, permet de confirmer que les deux premières techniques sont plus convaincantes pour des données dont le taux de variables indépendantes est faible, c'est-à-dire $s < 0.5$. En effet, le taux d'erreur augmente lorsque s augmente et que $s < 0.5$. Il diminue ensuite, une fois que $s > 0.5$. Pour la dernière technique, peu d'erreurs sont commises lorsque $s > 0.7$, ce qui est attendu car cette technique permet de construire des modèles graphiques possédant peu d'arêtes. Le taux d'erreur diminue donc, plus s augmente.

5.2 Taux d'erreur sous la loi normale asymétrique

Après avoir estimé le taux d'erreur des différentes techniques sur base de données suivant une loi multinormale, le cas de données suivant une autre loi est étudié. Comme déjà utilisée dans les chapitres 2 et 4, il s'agit de la loi normale asymétrique, car, pour celle-ci, les dépendances entre les variables sont connues à partir des paramètres α et Ω (voir section 2.3.2). Comme, au vu des exemples sur la loi normale, utiliser les méthodes AIC et BIC n'est pas adéquat pour trouver la valeur du paramètre de pénalisation, ce dernier sera choisi de sorte à minimiser le taux d'erreur.

Pour estimer le taux d'erreur des différentes techniques, $n = 200$ données sont générées selon la loi normale asymétrique de paramètres α et Ω. La matrice Y peut, dans ce cas, être définie par

$$Y_{ij} = \begin{cases} 0 & \text{si } \Omega^{-1}_{ij} = 0 \text{ et } \alpha_i \alpha_j = 0 \\ 1 & \text{sinon.} \end{cases}$$

Les taux de faux positif, de faux négatif et d'erreur peuvent donc être calculés sur base de ces deux paramètres et de la matrice d'adjacence du graphe estimé.

Tout comme dans la première partie de la section précédente, différentes structures pour les deux paramètres seront utilisées pour créer les différents exemples et les dimensions $d = 10$ et $d = 20$ seront étudiées. Ces structures sont aussi choisies de sorte à faire varier le taux s de couples de variables indépendantes. Voici la description de ces différents paramètres dans le cas de la dimension $d = 10$:

— Cas indépendant ($s = 1$) :
  - $\Omega^{-1} = I_{10}$
  - $\alpha = (0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$

— Cas équi-corrélé ($s = 0$) :
  - $\Omega^{-1}_{ij} = 0.4 \ \forall i \neq j$ et $\Omega^{-1}_{ii} = 1 \ \forall i$
  - $\alpha = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$

— Cas 1 ($s = 0.4$) :
  - $\Omega^{-1}_{i,i+1} = \Omega^{-1}_{i+1,i} = 0.2 \quad \forall i \in \{1, \cdots, 9\} \quad \Omega^{-1}_{i,i+2} = \Omega^{-1}_{i+2,i} = 0.2 \quad \forall i \in \{1, \cdots, 8\}$, $\Omega^{-1}_{i,j} = 0$ sinon et $\Omega^{-1}_{i,i} = 1 \quad \forall i \in \{1, \cdots, 10\}$

- $\alpha = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$

— Cas 2 ($s = 0.8$) :

- $\Omega^{-1}_{i,i+1} = \Omega^{-1}_{i+1,i} = 0.4 \quad \forall i \in \{1, \cdots, 6\}, \quad \Omega^{-1}_{i,j} = 0$ sinon et $\Omega^{-1}_{i,i} = 1 \quad \forall i \in \{1, \cdots, 10\}$
- $\alpha = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)$

— Cas 3 ($s = 0.2$) :

- $\Omega^{-1}_{i,i+1} = \Omega^{-1}_{i+1,i} = 0.2 \quad \forall i \in \{1, \cdots, 9\}, \quad \Omega^{-1}_{i,i+8} = \Omega^{-1}_{i+8,i} = 0.2 \quad \forall i \in \{1, 2\}, \quad \Omega^{-1}_{i+3,i+8} = \Omega^{-1}_{i+8,i+3} = 0.2 \quad \forall i \in \{1, 2\}, \quad \Omega^{-1}_{i+5,i+8} = \Omega^{-1}_{i+8,i+5} = 0.2 \; \forall i \in \{1, 2\}, \quad \Omega^{-1}_{i,j} = 0$ sinon et $\Omega^{-1}_{i,i} = 1 \quad \forall i \in \{1, \cdots, 10\}$
- $\alpha = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0)$

Les moyennes des taux d'erreur, de faux positif et de faux négatif pour ces exemples, basés sur $N = 100$ itérations sont reprises dans la Table 5.5.

$d = 10$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0	0	0
		NP	0	0	0
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.0002	0.0002	0
		NP	0.0008	0.0008	0
		arbre	0.9973	0.9973	0
Cas 1	0.4	*glasso*	0.2297	0.1108	0.1188
		NP	0.2322	0.1015	0.1306
		arbre	0.5922	0.5922	0
Cas 2	0.8	*glasso*	0.0593	0.0562	0.0031
		NP	0.0588	0.0555	0.0033
		arbre	0.0804	0.0775	0.0028
Cas 3	0.2	*glasso*	0.1806	0.0202	0.1604
		NP	0.1906	0.0257	0.1648
		arbre	0.7726	0.7717	0.0008

TABLE 5.5 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi normale asymétrique de dimension $d = 10$.

Au vu de la Table 5.5, le taux d'erreur est toujours plus élevé lorsque s possède une valeur proche de 0.4 pour les deux premières techniques. Ce taux est plus faible, pour la technique d'estimation sous forme d'arbre, lorsque $s = 0.8$ ou $s = 1$. Le plus interpellant dans cet exemple est que le taux d'erreur de la technique *non paranormale* est similaire à celui de la technique *glasso*. Or, pour cette dernière méthode, l'hypothèse de multinormalité des données est violée.

Lorsque la dimension des données est de $d = 20$, les paramètres de la loi normale asymétrique sont adaptés, mais les différentes valeurs du taux s de variables indépendantes sont toujours respectées. Voici la description de ces différents paramètres dans le cas de la dimension $d = 20$:

— Cas indépendant ($s = 1$) :
- $\Omega^{-1} = I_{20}$
- $\alpha_{20} = 1$ et $\alpha_i = 0 \quad \forall i \in \{1, \cdots, 19\}$

— Cas équi-corrélé ($s = 0$) :
- $\Omega_{ij}^{-1} = 0.4 \; \forall i \neq j$ et $\Omega_{ii}^{-1} = 1 \quad \forall i$
- $\alpha_i = 1 \quad \forall i \in \{1, \cdots, 10\}$ et $\alpha_i = 0 \; \forall i \in \{11, \cdots, 20\}$

— Cas 1 ($s = 0.4$) :
- $\Omega_{i,i+1}^{-1} = \Omega_{i+1,i}^{-1} = 0.2 \quad \forall i \in \{1, \cdots, 19\} \quad \Omega_{i,i+3}^{-1} = \Omega_{i+3,i}^{-1} = 0.2 \quad \forall i \in \{1, \cdots, 12\}$, $\Omega_{19,16}^{-1} = \Omega_{16,19}^{-1} = 0.2$, $\Omega_{1,17}^{-1} = \Omega_{17,1}^{-1} = 0.2$, $\Omega_{2,18}^{-1} = \Omega_{18,2}^{-1} = 0.2$, $\Omega_{3,19}^{-1} = \Omega_{19,3}^{-1} = 0.2$, $\Omega_{4,20}^{-1} = \Omega_{20,4}^{-1} = 0.2$, $\quad \Omega_{i,j}^{-1} = 0$ sinon et $\Omega_{i,i}^{-1} = 1 \quad \forall i \in \{1, \cdots, 10\}$
- $\alpha_i = 1 \; \forall i \in \{1, \cdots, 15\}$ et $\alpha_i = 0 \; \forall i \in \{16, \cdots, 20\}$

— Cas 2 ($s = 0.8$) :
- $\Omega_{i,i+1}^{-1} = \Omega_{i+1,i}^{-1} = 0.4 \quad \forall i \in \{1, \cdots, 19\}, \quad \Omega_{i,j}^{-1} = 0$ sinon et $\Omega_{i,i}^{-1} = 1 \quad \forall i \in \{1, \cdots, 10\}$
- $\alpha_i = 1 \; \forall i \in \{1, \cdots, 8\}$ et $\alpha_i = 0 \; \forall i \in \{9, \cdots, 20\}$

— Cas 3 ($s = 0.2$) :
- $\Omega_{i,i+1}^{-1} = \Omega_{i+1,i}^{-1} = 0.2 \quad \forall i \in \{1, \cdots, 19\}, \quad \Omega_{i,i+17}^{-1} = \Omega_{i+17,i}^{-1} = 0.2 \quad \forall i \in \{1, 2, 3\}, \quad \Omega_{i+3,i+17}^{-1} = \Omega_{i+17,i+3}^{-1} = 0.2 \quad \forall i \in \{1, 2, 3\}, \quad \Omega_{i+5,i+17}^{-1} = \Omega_{i+17,i+5}^{-1} = 0.2 \quad \forall i \in \{1, 2, 3\}, \quad \Omega_{i+6,i+17}^{-1} = \Omega_{i+17,i+6}^{-1} = 0.2 \quad \forall i \in \{1, 2, 3\}, \quad \Omega_{18,20}^{-1} = \Omega_{20,18}^{-1} = 0.2, \quad \Omega_{i,j}^{-1} = 0$ sinon et $\Omega_{i,i}^{-1} = 1 \quad \forall i \in \{1, \cdots, 10\}$
- $\alpha_i = 1 \; \forall i \in \{1, \cdots, 17\}$ et $\alpha_i = 0 \; \forall i \in \{18, \cdots, 20\}$

Les moyennes des taux d'erreur, de faux positif et de faux négatif de ces exemples, basés sur $N = 100$ itérations sont reprises dans le tableau 5.6.

Le fait que la technique *non paranormale* n'ait pas de meilleurs résultats que la technique *glasso* dans le cas de cet exemple est, en réalité, dû au fait que les données générées selon la loi normale asymétrique n'ont pas une distribution suffisamment éloignée de la distribution multinormale. En effectuant les tests de multinormalité de Mardia et de Henze-Zirkler sur des données générées selon la loi normale asymétrique, la multinormalité des données n'est rejetée pour aucun des deux tests. La distribution des données n'étant pas suffisamment éloignée de la distribution normale, utiliser les transformations de la technique *non paranormale* avant d'utiliser la technique *glasso* ne permet pas d'obtenir de meilleurs résultats.

$d = 20$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0.0251	0	0.0251
		NP	0.0250	0	0.0250
		arbre	0.0002	0	0.0002
Cas équi-corrélé	0	*glasso*	0	0	0
		NP	0	0	0
		arbre	0.9654	0.9654	0
Cas 1	0.4	*glasso*	0.3276	0.0021	0.3255
		NP	0.3338	0.0018	0.3320
		arbre	0.4413	0.4317	0.0095
Cas 2	0.8	*glasso*	0.3782	0	0.3782
		NP	0.3774	0	0.3774
		arbre	0.0008	0.0008	0
Cas 3	0.2	*glasso*	0.2216	0.0033	0.2183
		NP	0.2215	0.0023	0.2192
		arbre	0.5828	0.5803	0.0025

TABLE 5.6 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi normale asymétrique de dimension $d = 20$.

De plus, en comparant les Tables 5.3 et 5.5 entre elles ainsi que la Table 5.4 avec la Table 5.6, le taux d'erreur moyen pour cette loi est plus élevé que le taux d'erreur basé sur la loi multinormale. Cela vient du fait que les deux premières techniques ne tiennent pas compte du paramètre α intervenant dans la définition de la dépendance entre deux variables. Ainsi, la technique *glasso* effectue un certain nombre d'erreurs lorsque les données sont distribuées selon une loi qui n'est pas multinormale mais pas non plus trop éloignée de la multinormale. De plus, la technique *non paranormale* ne permet pas de diminuer ce taux d'erreur. Le taux d'erreur pour la technique sous forme d'arbre reste, quand à lui similaire, voire légèrement plus faible pour cette loi que pour la loi multinormale, en comparant les tableaux 5.1 avec 5.5 et 5.2 avec 5.6. Cette technique semble donc tenir compte des dépendances effectives reprises dans la base de données.

En conclusion, lorsque les données suivent une loi proche d'une loi multinormale mais qui n'en est pas une non plus, la technique *glasso* effectue un certain nombre d'erreurs car les données sont considérées comme étant multinormales. Ainsi, seules les dépendances reprises dans la matrice de concentration sont prises en compte dans la structure du graphe. Ensuite, la technique *non paranormale* ne permet pas d'obtenir de meilleurs résultats. Ces deux techniques ne semblent dont pas adaptées à ce type de données.

5.3 Transformation de la loi multinormale

A titre de second exemple non gaussien, plusieurs articles, comme [21] et [22], commencent par générer des données selon une loi multinormale et appliquent ensuite, à toutes les variables, une transformation. Cette transformation peut être de différents types. Dans ce mémoire, deux types de transformations seront utilisées. De plus, pour faciliter les simulations, ces transformations seront identiques pour toutes les variables. Lorsque des données sont générées de cette façon, le modèle graphique créé sur base des données multinormales de départ devrait être identique au modèle graphique obtenu sur les données après transformation. En effet, les transformations sont appliquées sur chaque variable indépendamment des autres variables, les dépendances entre les variables restent donc inchangées.

Pour ces deux cas, les mêmes structures de matrice de concentration que pour le cas normal sont utilisées. Les taux d'erreur, de faux positif et de faux négatif moyens seront alors calculés pour les 5 exemples en dimension 10 et ensuite pour les 5 exemples en dimension 20.

5.3.1 Transformation via la fonction de répartition

La première fonction de transformation est la fonction de répartition d'une loi normale de moyenne 0.2 et d'écart-type égal à 0.4. Les fonctions de densité marginales univariées deviennent alors bi-modales et plus normales, comme le montre la Figure 5.3 représentant les fonctions des densité estimées par noyau d'un vecteur univarié de 500 données, généré selon une loi normale centrée réduite, avant et après transformation.

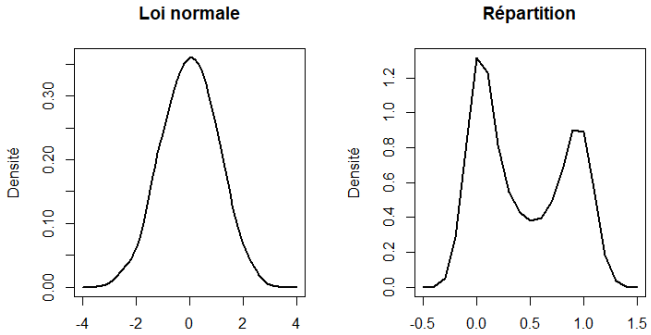

FIGURE 5.3 – Fonctions de densité estimées par noyau d'un vecteur univarié de 500 données, généré selon une loi normale centrée réduite, avant et après transformation à l'aide de la fonction de répartition d'une loi normale de moyenne 0.2 et d'écart-type égal à 0.4.

La Table 5.7 reprend les résultats des moyennes des taux d'erreur, de faux positif et de faux négatif calculés sur base de 100 simulations et dans le cas de $n = 200$ données générées en dimension $d = 10$.

$d = 10$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0	0	0
		NP	0	0	0
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.014	0.014	0
		NP	0	0	0
		arbre	1	1	0
Cas par bloc 1	0.4	*glasso*	0.2495	0.0557	0.1937
		NP	0.1788	0.0315	0.1473
		arbre	0.6222	0.6222	0
Cas par bloc 2	0.8	*glasso*	0.1068	0.0664	0.0404
		NP	0.0660	0.0317	0.0342
		arbre	0.2222	0.2222	0
Cas par bloc 3	0.2	*glasso*	0.1711	0.0144	0.1566
		NP	0.1222	0.0048	0.1173
		arbre	0.8000	0.8000	0

TABLE 5.7 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 10$, transformée à l'aide de la fonction de répartition.

La Table 5.8 contient les différents taux dans le cas où la dimension est égale à 20.

Lorsque les données sont générées de la façon précisée ci-dessus, celles-ci ne suivent pas une loi multinormale. Les tests de Mardia et de Henze-Zirkler confirment cela en rejetant la multinormalité. Il est donc attendu que la technique *non paranormale* ait un taux d'erreur inférieur à celui de la technique *glasso*, ce qui est bien le cas. Une discussion plus approfondie sur ces résultats sera donnée dans les sections 5.3.2 et 5.3.3.

$d = 20$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0	0	0
		NP	0	0	0
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.1035	0.1035	0
		NP	0.0268	0.0268	0
		arbre	1	1	0
Cas par bloc 1	0.4	*glasso*	0.3753	0.0685	0.3067
		NP	0.3114	0.0472	0.2641
		arbre	0.6052	0.6052	0
Cas par bloc 2	0.8	*glasso*	0.1842	0.1743	0.0098
		NP	0.1570	0.1335	0.0235
		arbre	0.1894	0.1894	0
Cas par bloc 3	0.2	*glasso*	0.2531	0.0887	0.1644
		NP	0.2092	0.0395	0.1696
		arbre	0.8052	0.8052	0

TABLE 5.8 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 20$, transformée à l'aide de la fonction de répartition.

5.3.2 Transformation via une fonction puissance

La deuxième fonction de transformation est la fonction puissance. Cette fonction a la forme

$$g(t) = \text{sign}(t)|t|^a,$$

où $a > 0$ et $\text{sign}(t) = 1$ si $t \geq 0$ et -1 sinon. Pour cette section, a est choisi comme étant égal à 3. Les fonctions de densité marginales univariées ont alors la forme d'un pic très aigu, comme le montre la Figure 5.4 représentant les fonctions des densité estimées par noyau d'un vecteur univarié de 500 données, généré selon une loi normale centrée réduite, avant et après transformation.

La Table 5.9 reprend les résultats des moyennes des taux d'erreur, de faux positif et de faux négatif calculés sur base de 100 simulations et dans le cas de $n = 200$ données générées en dimension $d = 10$.

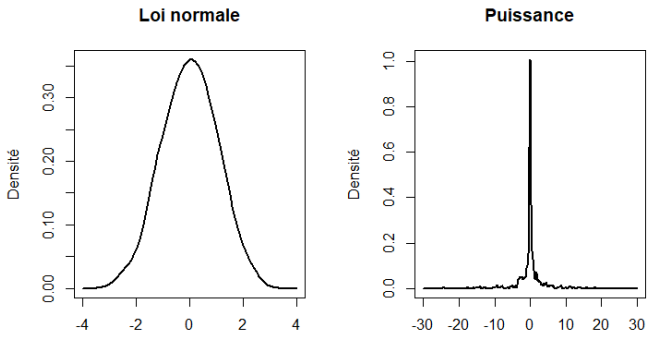

FIGURE 5.4 – Fonctions des densité estimées par noyau d'un vecteur univarié de 500 données, généré selon une loi normale centrée réduite, avant et après transformation à l'aide de la fonction puissance $g(t) = sign(t)|t|^3$.

$d = 10$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0	0	0
		NP	0	0	0
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.0393	0.0393	0
		NP	0	0	0
		arbre	1	1	0
Cas par bloc 1	0.4	*glasso*	0.2833	0.0715	0.2217
		NP	0.1824	0.0328	0.1495
		arbre	0.6222	0.6222	0
Cas par bloc 2	0.8	*glasso*	0.1626	0.1060	0.0566
		NP	0.0673	0.0311	0.0362
		arbre	0.2222	0.2222	0
Cas par bloc 3	0.2	*glasso*	0.1862	0.0320	0.1542
		NP	0.1284	0.0075	0.1208
		arbre	0.8000	0.8000	0

TABLE 5.9 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 10$, transformée à l'aide d'une fonction puissance.

La Table 5.10 contient les différents taux dans le cas où la dimension est égale à 20.

$d = 20$	s	Technique	Taux d'erreur	Taux de FN	Taux de FP
Cas indépendant	1	*glasso*	0	0	0
		NP	0	0	0
		arbre	0	0	0
Cas équi-corrélé	0	*glasso*	0.1010	0.1010	0
		NP	0.0260	0.0260	0
		arbre	1	1	0
Cas par bloc 1	0.4	*glasso*	0.3699	0.0732	0.2966
		NP	0.3082	0.0477	0.2604
		arbre	0.6052	0.6052	0
Cas par bloc 2	0.8	*glasso*	0.1830	0.1754	0.0075
		NP	0.1574	0.1321	0.0252
		arbre	0.1894	0.1894	0
Cas par bloc 3	0.2	*glasso*	0.2551	0.0887	0.1663
		NP	0.2067	0.0375	0.1692
		arbre	0.8052	0.8052	0

TABLE 5.10 – Moyennes des taux d'erreur, de faux négatif et de faux positif pour différents exemples générés à partir d'une loi multinormale de dimension $d = 20$, transformée à l'aide d'une fonction puissance.

Lorsque les données sont modifiées à l'aide d'une fonction puissance, la différence entre le taux d'erreur de la technique *glasso* et la technique *non paranormale* est plus élevée, surtout dans le cas de la dimension $d = 10$. Cette différence dépendra donc de la forme des fonctions de densité univariées et surtout de la mesure de différence entre la densité multivariée et la densité multinormale. Ainsi, il est clair, dans ce cas, que la technique *non paranormale* doit être privilégiée à la technique *glasso* pour estimer le graphe de données dont la multinormalité n'est pas respectée.

De plus, au vu des tableaux 5.3, 5.7 et 5.9, le taux d'erreur pour la technique *non paranormale* est toujours de la même grandeur au sein du même exemple. Cela montre donc bien que le graphe estimé à l'aide de cette technique sera identique qu'une transformation soit appliquée aux données multinormales ou non ; ce qui n'est pas le cas de la technique *glasso*.

Le taux d'erreur de la technique d'estimation sous forme d'arbre est lui aussi toujours identique lorsque la structure du graphe attendu est inchangée et ce, quelle que soit la fonction de densité des données. Le taux d'erreur de cette technique dépend uniquement du nombre de couples de variables dépendantes.

5.3.3 Comparaison graphique du taux d'erreur sous la loi normale et lorsqu'elle est transformée

Pour une plus grande facilité d'interprétation, la comparaison des taux d'erreur des 3 techniques lorsque les deux transformations ci-avant sont appliquées aux données suivant une loi multinormale peut se faire graphiquement. Pour ce faire, la technique de simulation développée dans la section 5.1.2 est reprise ; c'est-à-dire que le taux d'erreur sera étudié en fonction du taux de variables indépendantes. Des boites à moustaches représentant le taux d'erreur, de faux positif ainsi que de faux négatif seront représentées pour chaque technique et chaque exemple (loi normale et les deux transformations). Le nuage de points représentant le taux d'erreur en fonction du taux de variables indépendantes s sera également représenté pour les différents cas.

Pour effectuer cette comparaison, une matrice de concentration Θ est générée de sorte à avoir un taux s de valeurs nulles dans le triangle inférieur, avec $s \in [0, 1]$. Ensuite $n = 200$ données sont générées selon une loi $\mathcal{N}_{10}(0, \Theta^{-1})$. Pour considérer le cas normal, ces données sont conservées telles quelles. Pour obtenir des données non normales, les transformations utilisées précédemment sont appliquées à ces mêmes données. Cela permet donc de comparer les trois techniques sur des exemples identiques. Une fois que ces trois ensembles de données sont obtenus, les trois techniques d'estimation de structure de graphe sont appliquées sur chacun des ensembles. Au total, 9 taux d'erreur seront obtenus tout comme 9 taux de faux positif, de faux négatif et de valeurs nulles dans le triangle inférieur de la matrice de concentration. Cette procédure sera effectuée $N = 400$ fois pour pouvoir construire des graphiques suffisamment représentatifs des taux réels.

Les boites à moustaches représentées à la Figure 5.5 permettent de constater visuellement que le taux d'erreur de la technique *glasso* augmente lorsque les données ne suivent plus une loi normale. Cette augmentation est principalement due à l'augmentation du taux de faux négatif. Cela signifie que certains couples de variables sont considérés comme indépendants dans le modèle graphique alors qu'ils ne le sont pas en réalité. Le taux de faux positif n'augmente que légèrement dans le cas de la transformation par une fonction de puissance.

Concernant la technique d'estimation *non paranormale*, il est clair, au vu du deuxième graphique de la Figure 5.5, que le taux d'erreur reste inchangé, que les données suivent une loi normale ou non, lorsque la structure du graphe est identique. Les taux de faux positif et de faux négatif sont eux aussi égaux. Il est donc clair que la technique *non paranormale* permet de normaliser les données et que cette technique sera plus performante lorsque l'hypothèse de multinormalité est violée.

Pour la dernière technique, estimant les modèles graphiques sous forme d'arbre ou de forêt, il ressortait des Table 5.1, Table 5.7 et Table 5.9 que le taux d'erreur restait constant, que les données suivent une loi normale ou non. Cependant, au vu des boites à moustaches du dernier graphique de la Figure 5.5, celui-ci semble augmenter lorsque les données ont été transformées, et plus particulièrement, lorsque cette transformation est du type fonction

puissance. De plus cette augmentation de l'erreur est due à l'augmentation du taux de faux négatif, en tenant compte que le taux de faux positif a diminué. Ainsi, certains couples de variables sont considérés comme indépendants dans le modèle graphique alors qu'ils ne le sont pas en réalité. Cela pourrait peut-être être du au fait que la dépendance entre les variables est amoindrie lorsqu'une transformation est effectuée.

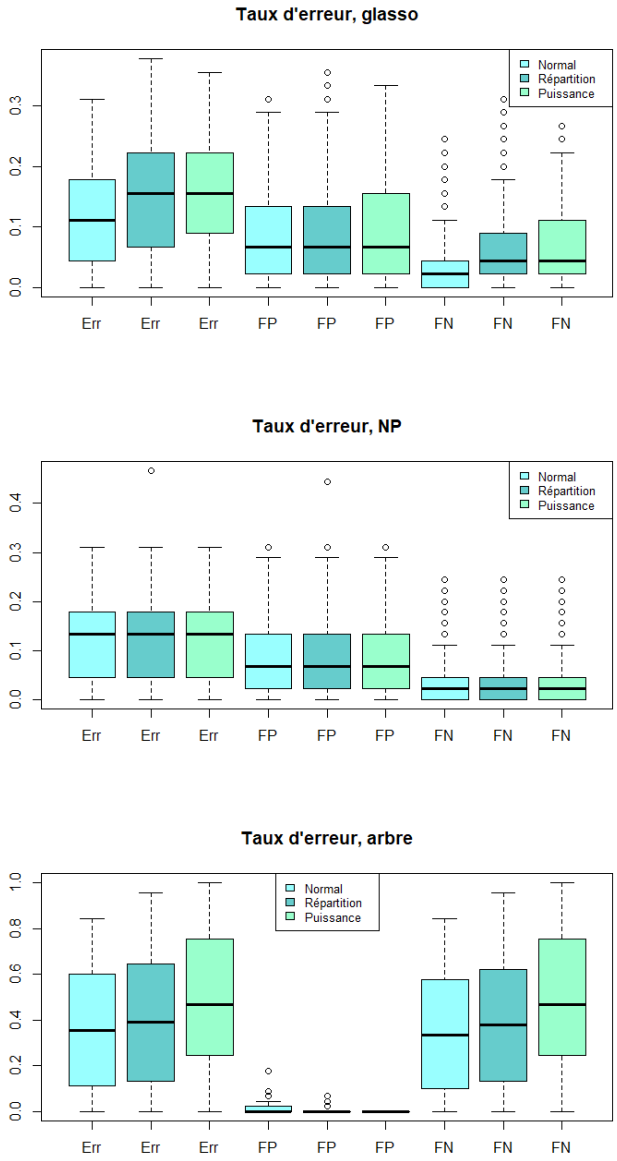

FIGURE 5.5 – Boites à moustaches représentant les taux d'erreur, de faux positif et de faux négatif pour les 3 techniques et selon que les données suivent une loi normale ou que la loi normale soit modifiée à l'aide de la fonction de répartition ou d'une fonction puissance.

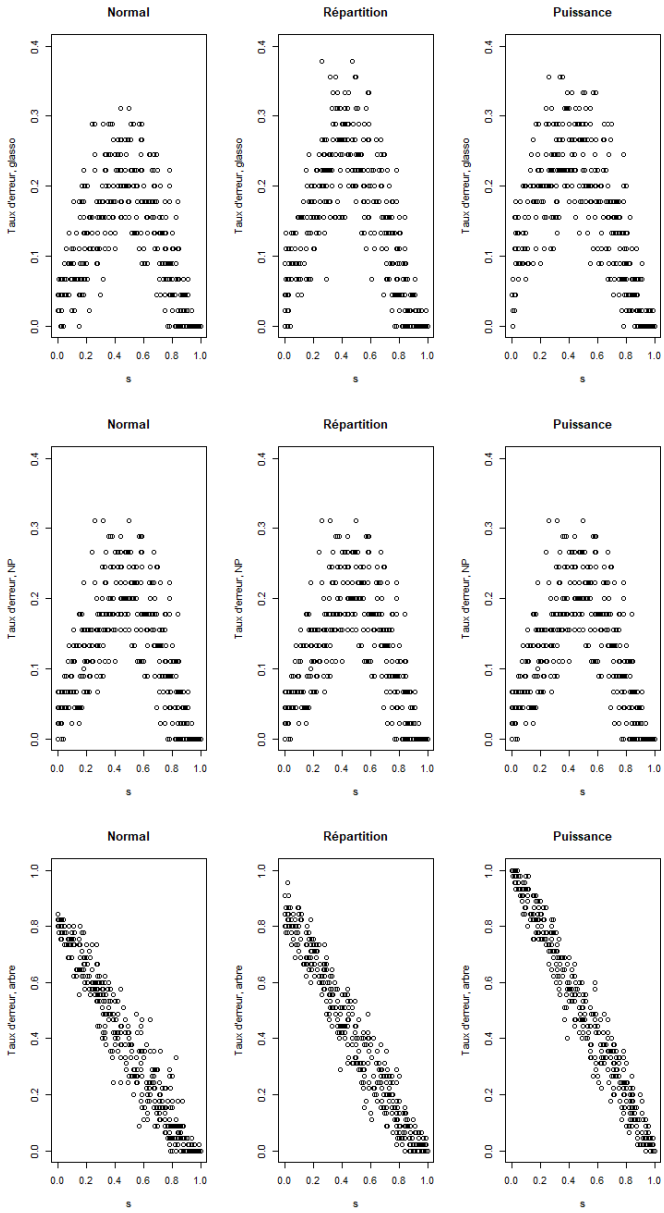

FIGURE 5.6 – Nuages de points représentant le taux d'erreur en fonction du taux de variables indépendantes pour les 3 techniques et selon que les données suivent une loi normale ou que la loi normale soit modifiée à l'aide de la fonction de répartition ou d'une fonction puissance.

Les nuages de points de la Figure 5.6 permettent de visualiser le taux d'erreur de chaque technique en fonction du taux de variables indépendantes et ce pour les trois types de données analysées. La première ligne de cette figure montre que le taux d'erreur pour la technique *glasso* augmente légèrement pour toutes les valeurs de s lorsque les données ne suivent plus une loi normale ; mais qu'une augmentation plus élevée a tout de même lieu lorsque s est proche de 0.4. Sur la deuxième ligne de la Figure 5.6, les nuages de points sont identiques, donc les modèles graphiques ont le même nombre d'erreurs que les données aient été transformées ou non, les modèles graphiques semblent donc être identiques. Finalement, les graphiques de la troisième ligne montrent que l'augmentation du taux d'erreur lorsque les données sont transformées est plus élevée lorsque s est proche de 0.

5.4 Conclusion

Lorsque les données suivent une loi multinormale, la technique *glasso* est à préférer. En effet, celle-ci possède le taux d'erreur le plus faible quel que soit le taux de variables indépendantes. Il est malgré tout nécessaire de trouver une bonne valeur pour le paramètre de pénalisation. Choisir cette valeur en minimisant le critère AIC ou BIC mène à un taux d'erreur plus grand que le taux d'erreur minimal possible. Ces critères permettent tout de même de créer des graphes possédant très peu de faux positifs. Ainsi, si deux variables ne sont pas reliées dans le graphe, celles-ci peuvent être supposées indépendantes avec un niveau de confiance assez grand.

Choisir le paramètre de pénalisation sans utiliser les critères AIC et BIC peut se faire de différentes manières. Dans ce mémoire, d'autres techniques ont aussi été utilisées. Si le graphe attendu doit posséder un nombre fixé d'arêtes, alors il est possible de choisir le paramètre de pénalisation de sorte à obtenir ce nombre d'arêtes. Si les dépendances à priori entre les variables de la base de données sont connues, alors le nombre de couples de variables dépendantes est un choix raisonnable pour le nombre d'arêtes. Si celles-ci ne sont pas connues, alors elles peuvent être en partie estimées sur base de la matrice de corrélation. Une autre façon de choisir le paramètre de pénalisation est de tracer le graphique du nombre d'arêtes du graphe en fonction de ce paramètre. Si ce graphique possède une zone constante à un certain nombre d'arêtes, alors ce nombre d'arêtes pourrait être un bon choix pour trouver la valeur du paramètre de pénalisation.

Lorsque les données ne suivent pas une loi multinormale, la technique *non paranormale* permettra d'obtenir de meilleurs résultats. Tout comme pour la technique précédente, il est nécessaire de trouver une valeur adéquate du paramètre de pénalisation.

La technique d'estimation sous forme d'arbre permet, quant à elle, d'estimer des modèles graphiques possédant peu d'arêtes. Lorsque la structure du modèle graphique est, à priori, proche de celle d'un arbre, c'est à dire que le graphe possède peu de cycles, alors le taux d'erreur de cette dernière méthode sera faible. Un autre avantage de cette technique est que le taux de faux positif est nul ou presque pour les différents exemples. Cela signifie

donc que lorsqu'une arête est présente dans la forêt créée sur base de cette technique, alors la dépendance entre ces variables existe réellement.

Pour obtenir un maximum d'informations sur les dépendances entre les variables d'une base de données, il peut donc être utile de combiner deux modèles graphiques :

1. Utiliser la technique *glasso* ou *non paranormale*, selon que les données suivent une loi multinormale ou non, pour définir des couples de variables indépendantes, sur base des couples de variables non liées dans le graphe.
2. Utiliser la technique d'estimation sous forme de forêt pour obtenir des couples de variables effectivement dépendantes.

Conclusion générale

Le but de ce mémoire était de représenter les dépendances entre les variables d'un vecteur aléatoire ou d'une base de données sous forme d'un modèle graphique. Pour ce faire, plusieurs techniques ont été développées, chacune ayant ses avantages et ses inconvénients.

Pour débuter, seul le cas des vecteurs multinormaux fut considéré. Pour ceux-ci, les dépendances entre les variables sont entièrement reprises dans la matrice de concentration. Il est donc nécessaire d'estimer cette matrice, mais pas seulement. En effet, lorsque cette dernière est estimée, elle ne possède aucun zéro, c'est-à-dire qu'aucun couple de variables n'est indépendant. Cela mène donc à un graphe complet et inutile dans l'étude des dépendances entre les variables. Pour contraindre les valeurs les plus faibles de l'estimation de la matrice de concentration à être nulles, la technique *graphical lasso* a été utilisée. Celle-ci consiste à estimer les valeurs de la matrice de concentration en maximisant une log-vraisemblance pénalisée. Qui dit pénalisation dit choix d'un paramètre de pénalisation. Un choix classique pour ce paramètre consiste à minimiser les critères AIC ou BIC. Cependant, il a été montré dans ce mémoire, que ces critères ne sont pas toujours adéquats.

Cette première technique étant basée sur l'hypothèse de multinormalité, une deuxième technique a été développée dans le but de pouvoir créer des modèles graphiques représentant les dépendances d'un vecteur aléatoire ayant une distribution continue quelconque. Cette seconde technique est nommée *non paranormale*. Celle-ci consiste à transformer chaque variable pour obtenir un vecteur aléatoire ayant une distribution multinormale et dont les dépendances entre les variables sont identiques aux dépendances comprises dans le vecteur aléatoire de départ. Lorsque le vecteur gaussien est obtenu, la technique *glasso* peut être appliquée afin d'obtenir une estimation de la structure du modèle graphique et de pouvoir le représenter.

Les modèles graphiques créés à partir des deux techniques ci-avant peuvent posséder un nombre élevé d'arêtes. Lorsque le but de l'analyse du vecteur aléatoire est de connaître les dépendances les plus fortes entre les variables, sans se soucier des dépendances qui pourraient être plus faibles, construire un modèle graphique possédant un nombre d'arêtes limité serait plus opportun. La troisième technique d'estimation de modèles graphiques présentée dans ce mémoire a pour but de restreindre la structure du graphe à un arbre ou une forêt. Celui-ci possèdera donc au maximum $d-1$ arêtes, où d est la dimension du vecteur aléatoire. Si davantage de couples de variables sont dépendants, alors seules les dépendances les plus fortes seront représentées dans le graphe.

Ces trois techniques ont finalement été comparées à travers leur taux d'erreur. Il ressort de cette comparaison que la technique d'estimation sous forme de forêt est à privilégier lorsque le taux de variables indépendantes est faible ou qu'un graphe possédant peu d'arêtes est demandé. De plus, cette technique peut être utilisée pour n'importe quel type de vecteur aléatoire dont les variables sont continues. Lorsqu'un graphe possédant un plus grand nombre d'arêtes est attendu, les techniques *glasso* et *non paranormale* peuvent être utilisées. Celles-ci se différencient par le type de données pour lesquelles elles sont les plus précises. La technique *glasso* est à privilégier pour les données gaussiennes, alors que la technique *non paranormale* est à privilégier pour les données ne vérifiant pas cette hypothèse.

Table des figures

Liste des tableaux

Bibliographie

[1] AZZALINI, A. et A. CAPITANIO. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 1999, 61(3), p. 579–602.

[2] BACH, Francis R et Michael I JORDAN. Beyond independent components : Trees and clusters. *Journal of Machine Learning Research*. 2004, 4(7-8), p. 1205–1233.

[3] BELILOVSKY, E., K. KASTNER, G. VAROQUAUX et M. BLASCHKO. Learning to discover probabilistic graphical model structures, 2016.

[4] BOYD, Stephen P. Convex optimization, 2004.

[5] CHAN, Terence H et Raymond W YEUNG. Probabilistic inference using function factorization and divergence minimization. *In : Towards an Information Theory of Complex Networks : Statistical Methods and Applications*, 1re éd., Boston : Birkhuser Boston, 2011, pages 47–74.

[6] EATON, Morris L. *Multivariate Statistics : A Vector Space Approach*. Place of publication not identified : Institute of Mathematical Statistics. ISBN 0-940600-69-2.

[7] FRIEDMAN, Jerome, Trevor HASTIE, Holger HÖFLING et Robert TIBSHIRANI. Pathwise coordinate optimization. *The Annals of Applied Statistics*. 2007, 1(2), p. 302–332.

[8] FRIEDMAN, Jerome, Trevor HASTIE et Robert TIBSHIRANI. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 200807, 9(3), p. 432,441.

[9] GÓMEZ, E, M.A GOMEZ-VIILEGAS et J.M MARÍN. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods*. 1998, 27(3), p. 589–600.

[10] HASTIE, Trevor. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* 2e éd. New York, NY : Springer New York : Imprint : Springer, 2009. (Springer Series in Statistics). ISBN 1-282-12674-1.

[11] HUBER, Peter J. *Robust statistics*. New York, NY : John Wiley. (Wiley series in probability and mathematical statistics). ISBN 0471418056.

[12] HØJSGAARD, Søren. *Graphical models with R.* 1re éd. New York : Springer, 2012. (Use R !). ISBN 1-280-80271-5.

[13] KANKAINEN, A., S. TASKINEN et H. OJA. On Mardia's tests of multinormality. *Mathematics Subject Classification*. 1991.

[14] KONISHI, Sadanori. *Information criteria and statistical modeling*. New York : Springer, 2008. (Springer series in statistics). ISBN 1-281-06732-6.

[15] KONISHI, Sadanori. Information criteria and statistical modeling, c2008.

[16] KRUSKAL, Joseph B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*. 1956-02-01, 7(1), p. 48,50.

[17] KULLBACK, Solomon. Information theory and statistics. *Journal of the Franklin Institute*. 1959, 268(1), p. 74–74.

[18] LAFFERTY, John, Han LIU et Larry WASSERMAN. Sparse nonparametric graphical models. *Statistical Science*. 2012, 27(4), p. 519–537.

[19] LAFIT, Ginette, Francisco J. NOGALES, Marcelo RUIZ et Ruben H. ZAMAR. A stepwise approach for high-dimensional gaussian graphical models. 2018.

[20] LIU, Han, Fang HAN, Ming YUAN, John LAFFERTY et Larry WASSERMAN. High-dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*. 2012, 40(4).

[21] LIU, Han, John LAFFERTY et Larry WASSERMAN. The nonparanormal : Semiparametric estimation of high dimensional undirected graphs. 2009.

[22] LIU, Han, Min XU, Haijie GU, Anupam GUPTA, John LAFFERTY et Larry WASSERMAN. Forest density estimation. *Carnegie Mellon University*. 2010.

[23] MEINSHAUSEN, N. et P. BÜHLMANN. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*. 2006, 34(3), p. 1436–1462.

[24] NELSEN, Roger B. *An Introduction to Copulas*. Vol. 139. Second edition éd. New York, NY : Springer New York, 2006. (Springer Series in Statistics). ISBN 9780387286594.

[25] SPEED, T. P. et H. T. KIIVERI. Gaussian markov distributions over finite graphs. *The Annals of Statistics*. 1986, 14(1), p. 138–150.

[26] TAN, Kean Ming, Daniela WITTEN et Ali SHOJAIE. The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational Statistics and Data Analysis*. mai 2015, 85, p. 23,36.

[27] TSYBAKOV, Alexandre B. *Introduction to Nonparametric Estimation*. 1re éd. New York, NY : Springer-Verlag, 2009. (Springer Series in Statistics). ISBN 9780387790510.

[28] VATS, Divyanshu, Robert D. NOWAK et Richard G. BARANIUK. Active learning for undirected graphical model selection, avril 2014.

[29] VOGEL, D. et R. FRIED. Elliptical graphical modelling. *Biometrika*. 2011, 98(4), p. 935–951.

[30] YUAN, Ming et Yi LIN. Model selection and estimation in the gaussian graphical model. *Biometrika*. 2007-03, 94(1), p. 19,35.

[31] ZAREIFARD, Hamid, Håvard RUE, Majid Jafari KHALEDI et Finn LINDGREN. A skew gaussian decomposable graphical model. *Journal of Multivariate Analysis*. 2016, 145, p. 58–72.

[32] ZHOU, Ming. A powerful test for multivariate normality. *Journal of Applied Statistics*. 2014-02-01, 41(2), p. 351,363.