UNIVERSITÉ LAVAL

La moyenne bayésienne pour les modèles basés sur les graphes acycliques orientés

Mémoire

Fatima Ezzahraa Bouzite

Maîtrise en statistique - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

La moyenne bayésienne pour les modèles basés sur les graphes acycliques orientés

Mémoire

Fatima ezzahraa Bouzite

Sous la direction de:

Denis Talbot, directeur de recherche
Geneviève Lefebvre, codirectrice de recherche

Résumé

Les méthodes d'inférence causale sont utiles pour répondre à plusieurs questions de recherche dans différents domaines, notamment en épidémiologie. Les graphes acycliques orientés sont des outils importants pour l'inférence causale. Entre autres, ils peuvent être utilisés pour identifier les variables confondantes utilisées dans l'ajustement de modèles statistiques afin d'estimer sans biais l'effet d'un traitement. Ces graphes sont construits à partir des connaissances du domaine d'application. Pourtant, ces connaissances sont parfois insuffisantes pour supposer que le graphe construit est correct. Souvent, un chercheur peut proposer divers graphiques correspondants à une même problématique.

Dans ce projet, on développe une alternative au modèle moyen bayésien traditionnel qui se base sur un ensemble de graphes proposés par un utilisateur. Pour sa mise en oeuvre, on estime d'abord la vraisemblance des données sous les modèles impliqués par chacun des graphes afin de déterminer la probabilité a posteriori de chaque graphe. On identifie, pour chaque graphe, un ensemble de covariables d'ajustement suffisant pour éviter le biais de confusion et on estime l'effet causal à partir d'approches appropriées en ajustant pour ces covariables. Finalement, l'effet causal global est estimé comme une moyenne pondérée des estimations correspondantes à chacun des graphes.

La performance de cette approche est étudiée à l'aide d'une étude de simulation où le mécanisme de génération des données est inspiré de l'étude Study of Osteoporotic

Fractures (SOF). Différents scénarios sont présentés selon les liens considérés entre les variables. L'étude de simulation démontre une bonne performance générale de notre méthode par comparaison au modèle moyen bayésien traditionnel.

L'application de cette approche est illustrée à l'aide de données de l'étude SOF dont l'objectif est l'estimation de l'effet de l'activité physique sur le risque de fractures de la hanche.

Abstract

Causal inference methods are useful for answering several research questions in different fields, including epidemiology. Directed acyclic graphs are important tools for causal inference.Among other things, they can be used to identify confounding variables used in fitting statistical models to unbiasedly estimate the effect of a treatment. These graphs are built from the knowledge of the domain of application. However, this knowledge is sometimes insufficient to assume that the constructed graph is correct. Often, a researcher can propose various graphs corresponding to the same problem.

In this project, we develop an alternative to the traditional Bayesian model averaging which is based on a set of graphs proposed by a user. For its implementation, we first estimate the likelihood of the data under the models implied by each graph to determine the posterior probability of each graph. A set of adjustment covariates sufficient to control for confounding bias is identified for each graph and the causal effect is estimated using appropriate approaches by adjusting for these covariates. Finally, the overall causal effect is estimated as a weighted average of the graph-specific estimates.

The performance of this approach is studied using a simulation study in which the data generation mechanism is inspired by the Study of Osteoporotic Fractures (SOF). Different scenarios varying in their relationships between the variables are presented.

The simulation study shows a good overall performance of our method compared to the traditional Bayesian model averaging.

The application of this approach is illustrated using data from the SOF, whose objective is to estimate the effect of physical activity on the risk of hip fractures.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Arrivé au terme de ce travail, il m'est particulièrement agréable d'exprimer mes remerciements à tous ceux qui, par leur enseignement, leur soutien et leurs conseils judicieux, m'ont permis de le mener à bien.

Je tiens tout d'abord à exprimer toute ma reconnaissance et mes remerciements les plus sincères à mon directeur de recherche Denis Talbot. Je le remercie pour sa grande disponibilité, son indispensable soutien ainsi que ses encouragements et son support moral dans les moments de doute ou de découragement. Ses conseils pertinents et ses révisions détaillées m'ont été d'une énorme utilité. Je tiens également à remercier ma codirectrice de recherche Geneviève Lefebvre pour sa précieuse collaboration, ses remarques et ses suggestions. Son attention aux détails et ses commentaires détaillés m'ont permet d'apporter plus de clarté et plus de cohérence à mon mémoire. Vos révisions, vos conseils ainsi que vos propositions ont grandement contribué à améliorer la qualité de mon travail. Veuillez trouver ici l'expression de ma plus profonde gratitude.

Un grand merci à mes chers parents. Que ce modeste travail soit l'exaucement de vos voeux tant formulés. Merci à mes frères Ahmed et Mohammed Reda et à ma soeur Meriem.

Je tiens à remercier mes chères amies Majida et Saida pour leurs encouragements, leur support moral et surtout leur capacité à relativiser les problèmes rencontrés.

Merci à ma chère amie Meriem pour sa gentillesse, sa positivité, ses encouragements et son sérieux dans tous les travaux que nous avons effectué ensemble. Merci pour tous ces beaux moments qu'on a vécus ensemble. Un immense merci à mes chères Asmaa, Oumaima, Sarah et Yasmina pour leur soutien, leur disponibilité et pour tous les beaux moments que nous avons partagés à l'université.

Je remercie également Awa, Miceline, David et Mamady pour leurs commentaires et leur disponibilité. Un merci aussi à Imane, Youssra, Hanaa, Chaimae, Ghizlane, Chaimaa et Fouad pour leurs conseils.

Introduction

L'inférence causale peut être définie comme l'ensemble des procédures qui tentent de prédire l'effet causal d'une variable exposition ou traitement [1] (par exemple, l'exposition à des matières dangereuses dans l'environnement, une intervention économique ou politique) sur une variable réponse ou issue (la santé ou les résultats socio-économiques). Les méthodes d'inférence causale se trouvent être un bon outil pour répondre à plusieurs questions de recherche dans différents domaines. Par exemple : la vaccination contre la COVID-19, l'un des sujets qui a suscité un immense intérêt au cours de ces derniers mois. En effet, en employant les méthodes d'inférence causale, les examinateurs scientifiques peuvent mieux prédire l'impact préventif qu'un vaccin pourrait avoir afin que les autorités puissent décider quant au budget de la campagne de vaccination.

La randomisation aboutit généralement à des inférences causales convaincantes (Hernán and Robins 2020). Toutefois, les études randomisées sont parfois difficiles à réaliser du fait de contraintes éthiques, économiques ou politiques. De ce fait, le recours à des études observationnelles semble être une alternative pertinente aux études avec assignation aléatoire du traitement. Or, il faut prendre en considération les facteurs de confusion d'une relation causale lors de l'analyse statistique de données observationnelles. Dans les expériences randomisées, le traitement est assigné d'une manière aléatoire, mais dans les études observationnelles, le traitement est

1. Les termes "exposition" et "traitement" sont utilisés de façon interchangeable dans ce mémoire.

généralement déterminé par de nombreux facteurs. Si ces facteurs sont associés à la fois au traitement et à la réponse, on dit alors qu'il y a une confusion. La confusion est souvent considérée comme la principale lacune des études observationnelles car sa présence peut donner une perception erronée de la relation de causalité en l'absence de contrôle.

La réduction ou l'élimination du biais de confusion en identifiant et en contrôlant pour les variables confondantes a généré un intérêt considérable auprès des statisticiens chercheurs. De nombreuses méthodes de sélection des facteurs de confusion ont été développées, mais l'utilisation de méthodes inappropriées peut produire des inférences inadéquates et nuire à la validité des résultats. De telles méthodes peuvent conduire, par exemple, à des estimateurs biaisés de l'effet causal de l'exposition si la méthode ne parvient pas à identifier les facteurs de confusion importants ou à des inférences imprécises si trop de covariables sont sélectionnées.

Les méthodes traditionnelles de sélection de variables comme le LASSO et le modèle moyen bayésien traditionnel se sont avérées plus efficaces que les procédures basées sur la signification statistique (la sélection descendante, la sélection ascendante, etc.) pour la construction de modèles de prédiction (Tibshirani 1996, Hoeting et al. 1999). Toutefois, elles ont été observées comme des méthodes médiocres pour la sélection des facteurs de confusion car elles font la modélisation de la variable réponse sans considérer le lien entre les facteurs et la variable d'exposition. Ainsi, des facteurs confondants importants associés fortement à l'exposition, mais faiblement à la réponse ont tendance à être négligés par ces approches. Par ailleurs, différents travaux ont montré qu'en plus des facteurs confondants, il est bénéfique d'inclure dans l'ensemble d'ajustement les prédicteurs purs de la réponse afin d'améliorer la précision des estimations (Shortreed and Ertefaie 2017, Talbot et al. 2015, Wang et al. 2012).

Plusieurs méthodes adaptées à la sélection des facteurs de confusion ont émergé ces dernières années. Ces méthodes prennent compte du fait que les facteurs de confusion

sont associés à la fois à l'exposition et à la réponse. Parmi ces méthodes émergentes, le LASSO adaptatif pour la réponse (Shortreed and Ertefaie 2017) qui asymptotiquement et sous certaines conditions, sélectionne le modèle de score de propension qui inclut tous les facteurs de confusion et les prédicteurs de la réponse, tout en excluant les autres covariables. Ainsi, il en résulte des estimations correctement ajustées pour la confusion et plus précises que celles d'un modèle entièrement ajusté. La question d'identification de variables confondantes a été aussi discutée dans Schneeweiss et al. 2009, Talbot et al. 2015, Rolling and Yang 2014 et Koch et al. 2018, notamment.

Or, les connaissances antérieures et le processus de génération de données fournissent des informations primordiales pour la sélection des variables confondantes [2]. Notamment, les graphes acycliques orientés (*Directed Acyclic Graphs* - DAG) sont reconnus pour faciliter la pensée causale. En résumant les connaissances d'experts et les hypothèses a priori sur la structure causale d'intérêt d'une manière intuitive, les DAG aident à clarifier les problèmes conceptuels et à améliorer la communication entre les chercheurs. Dans ces graphes, l'exposition, la réponse et les différentes variables pertinentes selon les connaissances du domaine d'application sont représentées et leurs relations présumées sont représentées par des flèches. Des règles peuvent alors être utilisées pour identifier les ensembles d'ajustement suffisants pour l'estimation de l'effet de l'exposition. Des règles ont même été récemment introduites pour identifier l'ensemble d'ajustement optimal parmi les ensembles suffisants (Rotnitzky and Smucler 2019, Henckel et al. 2019), c'est-à-dire l'ensemble suffisant permettant d'obtenir les estimateurs avec la plus petite variance asymptotique possible.

La construction des DAG se fait à l'aide des connaissances du domaine d'application, mais ces connaissances sont parfois insuffisantes pour supposer que le DAG construit est correct. Souvent, un chercheur peut imaginer plusieurs graphes compatibles avec ses connaissances. Dans le cadre de ce mémoire nous proposons une approche

2. Les termes "variables confondantes" et "facteurs de confusion" sont utilisés de manière interchangeable tout au long de ce mémoire.

d'analyse utilisant l'ensemble des graphes proposés par un utilisateur à partir du modèle moyen bayésien. Plus précisément, le projet nécessite d'abord d'estimer la vraisemblance des données sous les modèles impliqués par chacun des graphes afin de déterminer la probabilité a posteriori de chaque graphe. Parallèlement, pour chaque graphe, un ensemble de covariables d'ajustement suffisant pour éviter le biais de confusion est identifié et l'effet causal est estimé à partir d'approches appropriées en ajustant pour ces covariables. Finalement, l'effet causal global est estimé comme une moyenne des estimations correspondantes à chacun des graphes, où la moyenne est pondérée selon la probabilité a posteriori du graphe.

Le premier chapitre de ce mémoire est consacré à la présentation des outils et des concepts utilisés en inférence causale, ainsi qu'une introduction aux graphes acycliques orientés. Le deuxième chapitre présente une introduction à la statistique bayésienne, en particulier le modèle moyen bayésien et ses applications. Le troisième chapitre fournit une explication détaillée de la méthode du modèle moyen bayésien basé sur les DAG ainsi qu'une étude de simulation de type Monte-Carlo pour évaluer sa performance par comparaison au modèle moyen bayésien et à un modèle de régression brut. Le dernier chapitre illustre l'application de l'approche du modèle moyen bayésien basé sur les DAG sur les données de l'étude Study of Osteoporotic Fractures (SOF) pour estimer l'effet causal de la pratique de l'activité physique sur le risque de fractures de la hanche.

Chapitre 1

Introduction à l'inférence causale

1.1 Une définition de l'effet causal

1.1.1 L'effet causal individuel

Pour définir l'effet causal individuel, nous présentons l'exemple suivant : supposons qu'on a deux patients, Alain et Barry, qui sont atteints de la même maladie. À une date donnée, les deux patients reçoivent le même médicament. Une semaine plus tard, Alain meurt alors que Barry reste en vie. Imaginons que nous puissions savoir d'une manière ou d'une autre que si Alain n'avait pas reçu le médicament à cette date, il serait vivant une semaine plus tard et que si Barry n'avait pas reçu le médicament à cette date, il serait toujours en vie une semaine plus tard. En se basant sur ces informations, la plupart conviendrait que le médicament a causé la mort d'Alain. Par conséquent, le médicament a eu un effet causal sur la survie d'Alain à une semaine et il n'a pas eu d'effet causal sur la survie de Barry à une semaine. Pour rendre notre intuition causale accessible à l'analyse mathématique et statistique, nous allons introduire une notation (Rubin 1974). La notation définie ici garde des similitudes avec celle définie pour présenter les travaux de Hernán and Robins 2020.

Soit Y, la variable aléatoire correspondant à la réponse étudiée, dans notre exemple c'est une variable binaire (0 : survie, 1 : décès), et soit A, la variable aléatoire correspondant à l'exposition étudiée, dans notre exemple c'est une variable binaire (0 : non exposé, 1 : exposé). Afin de simplifier la présentation, nous supposons que Y et A sont des variables binaires, mais les concepts peuvent être généralisés pour d'autres types de Y et A. Nous notons également par l'indice $i = 1, ..., n$ les individus et utilisons les lettres minuscules y et a pour représenter les réalisations des variables aléatoires Y et A, respectivement. La variable $Y_i^{a=1}$ (respectivement $Y_i^{a=0}$) est la valeur de la réponse qui aurait été observée pour l'individu i sous la valeur de l'exposition $a = 1$ (respectivement $a = 0$). Les variables $Y_i^{a=1}$ et $Y_i^{a=0}$ sont appelées des réponses (issues) potentielles ou contrefactuelles.

Ainsi, une définition formelle d'un effet causal individuel est comme suit :

$$Y_i^{a=1} - Y_i^{a=0}.$$

On dit que l'exposition A a un effet causal sur la réponse Y d'un individu i si :

$$Y_i^{a=1} - Y_i^{a=0} \neq 0.$$

1.1.2 L'effet causal moyen

On dit que l'exposition A a un effet causal moyen non nul sur la réponse Y dans notre population d'intérêt si :

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \neq 0,$$

où, $\mathbb{E}[Y^a]$ est la réponse contrefactuelle moyenne si tous les individus de la population avaient reçu le niveau d'exposition a.

L'espérance mathématique des variables discrètes binaires est définie comme suit :

$$\mathbb{E}[Y^a] = \sum_{y \in \{0,1\}} yP[Y^a = y] = P[Y^a = 1].$$

Donc pour les réponses binaires, on dit qu'un effet causal moyen de l'exposition A sur la réponse Y est présent dans la population d'intérêt si :

$$P[Y^{a=1} = 1] - P[Y^{a=0} = 1] \neq 0.$$

Pour clarifier cette définition, reprenons notre exemple avec n patients. Supposons que les réponses de tous ces patients lorsqu'ils prennent le médicament ($a = 1$) et lorsqu'ils ne le prennent pas ($a = 0$) sont connues. Pour chaque patient, on a donc les valeurs $Y^{a=1}$ et $Y^{a=0}$. Si la différence des proportions de survie entre le cas où tous les patients auraient pris le médicament et le cas contraire est non nulle, alors le médicament a un effet causal moyen sur la maladie dans notre population d'intérêt.

1.2 Estimation de l'effet causal

1.2.1 Études randomisées

Le principe d'ignorabilité

En réalité, nous ne connaissons pas toutes les réponses contrefactuelles $Y^{a=1}$ sous traitement et $Y^{a=0}$ sans traitement. Au contraire, nous ne pouvons connaître que la réponse réellement observée Y sous le niveau de traitement effectivement reçu A. Ainsi, les données sont manquantes pour les autres réponses contrefactuelles dont nous avons besoin pour estimer l'effet causal.

La randomisation garantit que ces valeurs manquantes sont survenues d'une manière complètement aléatoire. Par conséquent, l'effet causal moyen peut être estimé dans les expériences randomisées malgré les données manquantes.

Pour éclaircir cette intuition, supposons qu'on a n patients séparés en deux groupes (g, g') et que l'assignation aux groupes est randomisée. Supposons que le groupe g reçoit le médicament ($A = 1$) et que le groupe g' ne le reçoit pas ($A = 0$). Une semaine après, on calcule le risque de mortalité dans chaque groupe : $P[Y = 1|A = 1]$ et $P[Y =$

$1|A = 0]$. Imaginons maintenant que le groupe g ne reçoit pas le médicament ($A = 0$) et que le groupe g' le reçoit ($A = 1$). Les proportions $P[Y = 1|A = 1]$ et $P[Y = 1|A = 0]$ seront les mêmes en absence de variabilité aléatoire car les sujets ont été assignés d'une façon aléatoire aux groupes. C'est ce qu'on appelle le principe *d'ignorabilité*, qui signifie que les résultats qu'on observe pour le groupe traité g auraient été les mêmes si le groupe g' avait été traité, et vice versa. Mathématiquement :

$$Y^a \perp A \quad \forall a \in \{0, 1\},$$

où : $\perp$ désigne le symbole d'indépendance statistique.

Pour que les réponses contrefactuelles d'un individu i, $Y_i^{a=1}$ et $Y_i^{a=0}$, aient un sens, il est nécessaire de faire les hypothèses de positivité, de stabilité et de cohérence décrites ci-dessous.

L'hypothèse de positivité

L'hypothèse de positivité se définit comme suit :

$$0 < P[A = a] < 1 \quad \forall a \in \{0, 1\}.$$

Ainsi chaque individu i a des probabilités non nulles d'être exposé et d'être non exposé, afin que Y_i^a soit bien identifié.

L'hypothèse de stabilité

La SUTVA (stable unit treatment value assumption) est une hypothèse nécessaire afin de définir la réponse potentielle. Cette hypothèse de stabilité comporte deux éléments, l'absence d'interférence et la version unique du traitement. Le premier élément stipule que la réponse contrefactuelle d'un individu sous la valeur d'exposition $A = a$ est indépendante des valeurs d'exposition des autres individus. L'interférence entre les sujets est courante dans le contexte des maladies infectieuses ou les programmes éducatifs où la réponse d'un individu est influencée par son interaction sociale avec

d'autres membres de la population. Le deuxième élément suppose qu'il n'y a pas de versions différentes d'un même traitement. Par exemple, supposons qu'on a deux médicaments semblables $M1$ et $M2$ et qu'on étudie l'effet de prendre l'un ou l'autre ($A = 1$ si une personne prend $M1$ ou $M2$ et $A = 0$ sinon). Imaginons qu'Alain pourrait rester en vie en prenant le médicament $M1$ et mourir en prenant le médicament $M2$. Dans ce cas la réponse contrefactuelle pour un individu n'est pas bien définie, car elle dépend de la version du traitement (le médicament utilisé).

L'hypothèse de cohérence

Pour chaque individu, l'une des réponses contrefactuelles - celle qui correspond à la valeur du traitement que l'individu a réellement reçu - est en fait factuelle. Autrement dit, un individu i avec un traitement observé $A_i = a$, a une réponse observée $Y_i = Y_i^a$; cette égalité est appelée cohérence.

Identifiabilité

Sous ces hypothèses, $\mathbb{E}[Y^{a=1}]$ et $\mathbb{E}[Y^{a=0}]$ peuvent être estimés sans biais par : $\bar{Y_1} = \frac{\sum_{i:a_i=1} Y_i}{n_1}$ et $\bar{Y_0} = \frac{\sum_{i:a_i=0} Y_i}{n_0}$ respectivement (Rubin 1974), où : n_0 et n_1 sont le nombre d'observations dans le groupe traité et le groupe non traité, respectivement, et $(i : a_i = a)$ désigne l'ensemble des i tels que $a_i = a$.

Démonstration :

$$\begin{aligned}
\mathbb{E}[Y^1] - \mathbb{E}[Y^0] &= \mathbb{E}[Y^1|A = 1] - \mathbb{E}[Y^0|A = 0] \quad (Y^a \perp A) \\
&= \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0] \quad (\text{par hypothèse de cohérence}) \\
&= \mathbb{E}\left[\frac{\sum_{i:a_i=1} Y_i}{n_1}\right] - \mathbb{E}\left[\frac{\sum_{i:a_i=0} Y_i}{n_0}\right] \\
&= \mathbb{E}[\bar{Y_1}] - \mathbb{E}[\bar{Y_0}].
\end{aligned}$$

1.2.2 Études observationnelles

L'hypothèse faible d'ignorabilité conditionnelle

Tout en reconnaissant que les expériences randomisées ont des avantages intrinsèques pour l'inférence causale, la réalisation de ce genre d'expérience en pratique n'est pas toujours possible à cause de contraintes éthiques, logistiques ou financières. C'est la raison pour laquelle les études observationnelles ont un rôle important dans la réponse à plusieurs questions de l'inférence causale.

Dans une étude observationnelle, différents facteurs peuvent agir à la fois sur l'exposition et la réponse observée. Ainsi, en présence de biais de confusion, l'estimation de l'effet causal moyen de l'exposition sera biaisée : $\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[\bar{Y}_1]$ et $\mathbb{E}[Y^{a=0}] \neq \mathbb{E}[\bar{Y}_0]$.

Pour estimer sans biais l'effet causal, nous pouvons contrôler pour un ensemble de variables confondantes L satisfaisant l'hypothèse faible d'ignorabilité conditionnelle (Rosenbaum and Rubin 1983) :

$$Y^a \perp A|L \quad \forall a \in \{0,1\}.$$

Ainsi, l'exposition A est indépendante de la réponse contrefactuelle dans les strates de L. Chaque strate est considérée comme une sous-popoulation dans laquelle l'ignorabilité est vérifiée.

L'hypothèse de positivité

Afin de pouvoir estimer l'effet causal, dans le cadre des études observationnelles, il faudrait que l'hypothèse de positivité soit vérifiée dans chacune des différentes strates formées par les combinaisons des variables confondantes. Mathématiquement :

$$0 < P[A = a|L = l] < 1 \quad \forall a \in \{0,1\}, \forall l.$$

Cette hypothèse ne serait pas respectée, par exemple, dans un contexte où l'on chercherait à comparer deux traitements, mais qu'un des deux est contrindiqué pour

certains patients en raison d'autres problèmes de santé. De tels patients auraient alors une probabilité nulle d'avoir le médicament contrindiqué.

Identification conditionnelle

Ces hypothèses ainsi que les hypothèses de stabilité et cohérence assurent l'identifiabilité des effets causaux à partir des données d'observation. Ainsi, l'effet causal moyen peut se calculer comme suit :

$$\begin{aligned}
\mathbb{E}[Y^1] - \mathbb{E}[Y^0] &= \mathbb{E}_L\{\mathbb{E}[Y^1|L=l]\} - \mathbb{E}_L\{\mathbb{E}[Y^0|L=l]\} \quad \text{(théorème de l'espérance totale)} \\
&= \mathbb{E}_L\{\mathbb{E}\left[Y^1|A=1,L=l\right]\} - \mathbb{E}_L\{\mathbb{E}[Y^0|A=0,L=l]\} \quad (Y^a \perp A|L) \\
&= \mathbb{E}_L\{\mathbb{E}[Y|A=1,L=l]\} - \mathbb{E}_L\{\mathbb{E}[Y|A=0,L=l]\} \\
&= \sum_l \left[\mathbb{E}[Y|A=1,L=l] - \mathbb{E}[Y|A=0,L=l]\right] P(L=l).
\end{aligned}$$

1.3 Introduction aux graphes acycliques orientés

L'utilisation des DAG est de plus en plus fréquente en épidémiologie ces dernières années suite aux travaux de Pearl, Robins, Greenland et autres (Greenland et al. 1999a, Glymour and Greenland 2008, Pearl 2009).

1.3.1 Description d'un DAG

Un DAG est un outil graphique qui permet de visualiser les relations entre l'exposition d'intérêt A, la réponse étudiée Y, ainsi que toutes les causes communes (même non mesurées) de deux variables quelconques du graphe. Les DAG comprennent donc un ensemble de variables (***nœuds/sommets***) avec des flèches dessinées entre elles pour montrer les directions des relations causales présumées. Cependant, aucune autre hypothèse sur la nature de ces relations n'est faite (par exemple, si l'exposition augmente ou diminue la valeur de la réponse). Par conséquent, l'absence de flèche

entre une paire de variables représente l'hypothèse qu'il n'y a pas de relation directe entre elles.

Le DAG est un graphe ***orienté*** et ***acyclique***. C'est-à-dire que dans un DAG, toutes les flèches ont une et une seule direction et en partant de n'importe quelle variable du graphe tout en suivant le sens des flèches, il est impossible de revenir à la variable de départ. Ce dernier point est justifié par le fait que toutes les causes précèdent temporellement les effets, et par conséquent, aucune variable ne peut être à la fois la cause et l'effet de toute autre variable du graphique.

1.3.2 Vocabulaire

Afin de faciliter la présentation de la suite de cette partie, nous présentons quelques notions utilisées dans les DAG.

Une cause est une variable qui influence, directement ou indirectement, la valeur d'une autre variable. Les causes sont souvent appelées ***ancêtres*** de l'autre variable, les causes directes sont appelées ***parents***. Un effet est une variable qui est influencée, soit directement ou indirectement, par une autre variable. Un effet est souvent appelé ***descendant***, un effet direct est appelé ***enfant***.

Un ***chemin*** (ou une ***voie***) fait référence à toute séquence de flèches qui relie deux variables ou plus, quelle que soit la direction des flèches. Un chemin ***orienté*** est un chemin dans lequel toutes les flèches se dirigent vers la même direction. Un chemin ***causal*** est un chemin orienté de l'exposition vers la réponse.

Soit un DAG où A est la variable exposition, Y est la variable réponse et C est une covariable associée d'une manière ou d'une autre à l'exposition et à la réponse, il existe trois ensembles de relations possibles :

1. Le premier scénario ($A \leftarrow C \rightarrow Y$) est que la covariable C est une cause commune de l'exposition et de la réponse. On dit que C est une variable ***confondante***. Le chemin entre les deux variables A et Y dans cet exemple est généralement appelé un chemin ***porte-arrière***.

2. Le deuxième scénario ($A \rightarrow C \rightarrow Y$) est que la covariable est sur la voie causale de l'exposition à la réponse. Dans ce cas, C est appelée une variable ***intermédiaire***.

3. La situation finale ($A \rightarrow C \leftarrow Y$) est celle où la covariable est un effet à la fois de l'exposition et de la réponse. Dans ce cas, C est appelée un ***collisionneur***.

1.3.3 DAG : Associations et biais de confusion

Un chemin ***ouvert*** dans un DAG est un chemin le long duquel une association peut être transmise. Toutes les voies non causales qui ne contiennent pas des collisionneurs sont ouvertes et génèrent un biais de confusion. En revanche, si un chemin est ***fermé*** ou ***bloqué***, aucune association statistique ne peut être générée le long de celui-ci. En absence de conditionnement sur le collisionneur ou sur un de ses descendants, toute voie non causale qui contient un collisionneur est fermée et ne génère pas de biais de confusion. Un chemin est également fermé si on contrôle pour une variable faisant partie du chemin qui n'est pas un collisionneur. Si tous les chemins entre deux variables sont bloqués, on dit qu'elles sont ***d-séparées***.

Une association entre deux variables nécessite au moins un chemin ouvert entre les variables. L'état des chemins peut être modifié par conditionnement. Un chemin ouvert peut être bloqué et vice versa en conditionnant sur des variables particulières le long de ce chemin. Notons que le conditionnement fait référence à un ajustement statistique, par exemple une analyse de régression, ou une stratification.

Exemples :

- Le chemin non causal ($A \leftarrow C \rightarrow Y$) est ouvert, donc il génère un biais de confusion. En conditionnant sur C, le chemin sera bloqué. Par conséquent, le contrôle d'une variable confondante ferme le chemin et supprime la confusion générée par ce chemin.

- Le chemin ($A \rightarrow C \rightarrow Y$) est ouvert. En supposant que le but est d'estimer l'effet total de A sur Y, ce chemin devrait être maintenu ouvert. Le conditionnement sur C dans cette situation bloquerait ce chemin et entraînerait une estimation de l'effet direct de A sur Y uniquement (c'est-à-dire une estimation de la partie qui n'est pas médiée par C).

- Le chemin ($A \rightarrow C \leftarrow Y$) est déjà bloqué. Cependant, si on contrôle le collisionneur C, alors le chemin sera ouvert et il introduira un biais de sélection qui n'existait pas précédemment.

1.3.4 Le critère porte-arrière

Afin d'estimer l'effet causal de A sur Y de façon non biaisée, il faut éliminer la confusion en contrôlant pour les variables confondantes. Pour obtenir un contrôle adéquat, ***le critère porte-arrière*** constitue une approche qui permet l'identification d'un ou de plusieurs ensembles d'ajustement suffisants (Pearl 2009).

Définition : Un ensemble de variables S satisfait le critère porte-arrière pour une paire de variables (A,Y) dans un DAG si :

1. aucun noeud de S n'est un descendant de A et
2. S bloque tous les chemins portes-arrières entre A et Y.

Exemple :

Dans la figure 1.1, le chemin porte-arrière ($A \leftarrow C \rightarrow Y$) génère un biais de confusion. Cependant, en contrôlant pour C, et en évitant de contrôler pour L, on obtient un ensemble d'ajustement suffisant $S = \{C\}$ qui permet d'éviter la confusion.

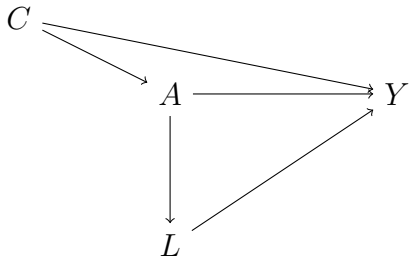


FIGURE 1.1 – Exemple d'un DAG

Or, des règles ont été récemment introduites pour identifier l'ensemble d'ajustement optimal parmi les ensembles suffisants (Rotnitzky and Smucler 2019, Henckel et al. 2019). En effet, un ensemble optimal devrait inclure les purs prédicteurs de la réponse (les variables qui ont des flèches vers Y, mais pas vers A), d'autant que ce ne soit pas des variables intermédiaires dans la relation entre A et Y. Cet ensemble ne devrait pas inclure les variables dites *instruments*, c'est-à-dire les variables avec une flèche vers A, mais pas vers Y. De plus, un ensemble d'ajustement optimal devrait satisfaire le critère porte-arrière. S'il est possible de bloquer un même chemin porte-arrière en ajustant pour différentes variables, il est préférable de prendre les variables les plus proches de Y sur le chemin causal. Par exemple, si on a ($A \leftarrow L_1 \rightarrow L_2 \rightarrow Y$), il est préférable de contrôler pour $\{L_2\}$ plutôt que $\{L_1\}$ ou $\{L_1, L_2\}$.

Chapitre 2

La statistique bayésienne : concepts et définitions

2.1 Introduction à la statistique bayésienne

2.1.1 Principes de base

La statistique bayésienne est un ensemble de méthodes qui partent des croyances a priori existantes sur un paramètre d'intérêt et les mettent à jour en utilisant des données observées pour donner des croyances a posteriori, qui fournissent la base pour des décisions inférentielles concernant le paramètre (Carlin and Louis 2008, Gelman et al. 2004, Duchesne 2012). Il y a trois éléments essentiels à l'approche bayésienne. Le premier élément est les connaissances de base sur un paramètre donné, θ, dans le modèle statistique considéré. Cet élément fait référence à toutes les connaissances disponibles avant la collecte des données et est capturé dans la loi de probabilité a priori, $\pi(\theta)$, qui permet de quantifier les incertitudes sur la valeur de θ. Le deuxième élément est l'information contenue dans les données observées y. Il s'agit de la vraisemblance des données, $\pi(y|\theta)$. Le troisième élément est basé sur la combinaison des deux premiers en utilisant le théorème de Bayes sous la forme de la

loi a posteriori $\pi(\theta|y)$ qui s'écrit comme suit :

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\int \pi(y|\theta)\pi(\theta)\, d\theta}. \tag{2.1}$$

La distribution a posteriori reflète la mise à jour des connaissances, en équilibrant les connaissances a priori avec les données observées, et est utilisée pour conduire les inférences.

La loi a priori peut dépendre d'autres paramètres, ϕ, qu'on appelle des hyper-paramètres. Ces derniers caractérisent le paramètre d'intérêt θ et donc la loi a priori est désignée par $\pi(\theta|\phi)$. Ainsi la loi a posteriori de θ se calcule en appliquant la règle des probabilités totales au numérateur et au dénominateur de l'équation (2.1). En supposant que l'indépendance conditionnelle : $\pi(y|\theta) = \pi(y|\phi, \theta)$ est vérifiée, on obtient :

$$\pi(\theta|y) = \frac{\int \pi(y|\theta)\pi(\theta|\phi)\pi(\phi)\, d\phi}{\iint \pi(y|\theta)\pi(\theta|\phi)\pi(\phi)\, d\phi\, d\theta}. \tag{2.2}$$

2.1.2 Les lois a priori non informatives

Le choix des lois a priori est une étape fondamentale dans l'analyse bayésienne. Une loi a priori non informative peut être définie comme une loi qui ne contient aucune information sur θ ou encore une loi où toutes les valeurs possibles de θ sont équiprobables. Supposons que l'ensemble des valeurs possibles de θ est discret et fini de taille q, une loi a priori non informative pourra être une loi qui assigne la probabilité $\frac{1}{q}$ à chaque valeur possible de θ. Or, cette approche soulève un problème très important : elle est non invariante par reparamétrisation. Par exemple, si θ a comme loi a priori la loi uniforme sur $[0,1]$: $\pi_\theta(\theta) = \mathcal{U}(\theta|0,1)$ et si $\phi = \ln(\frac{\theta}{1-\theta})$ est une

reparamétrisation du modèle, alors la loi a priori de ϕ est :

$$\begin{aligned}\pi_\phi(\phi) &= \pi_\theta\left(\frac{\exp(\phi)}{1+\exp(\phi)}\right)\left|\frac{\partial}{\partial\phi}\,\frac{\exp(\phi)}{1+\exp(\phi)}\right| \\ &= I\left[0<\frac{\exp(\phi)}{1+\exp(\phi)}<1\right]\,\frac{\exp(\phi)}{(1+\exp(\phi))^2} \\ &= \frac{\exp(\phi)}{(1+\exp(\phi))^2},\ \phi\in\mathbb{R},\end{aligned}$$

La densité π_ϕ est informative dans le sens où chaque valeur de ϕ dans $\mathbb{R}$ n'est pas aussi plausible qu'une autre.

La loi a priori de Jeffreys, fondée sur l'information de Fisher $I(\theta)$, est un exemple de loi non informative invariante par reparamétrisation de la loi a priori :

$$\pi_\theta(\theta) \propto \sqrt{|I(\theta)|},$$

où :

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2\ \ln(\pi(y|\theta))}{\partial\theta^2}\ |\theta\right].$$

2.2 L'approche du modèle moyen bayésien

Le modèle moyen bayésien (*Bayesian Model Averaging* - BMA) (Raftery et al. 1997, Hoeting et al. 1999) est une extension des méthodes habituelles d'inférence bayésienne dans lesquelles on ne modélise pas seulement l'incertitude du paramètre par la loi a priori, mais aussi l'incertitude du modèle. Le BMA implique d'obtenir les lois a posteriori des paramètres et des modèles en utilisant le théorème de Bayes et permet ainsi des performances prédictives améliorées par comparaison aux inférences basées sur la pratique statistique standard.

Soient $\mathcal{M} = (M_1, ..., M_K)$, l'ensemble des modèles considérés, θ_l pour $l \in [1, ...,K]$, le vecteur des paramètres du modèle M_l, $\pi(\theta_l|M_l)$, la loi a priori de θ_l sous le modèle M_l, et $\pi(y|\theta_l, M_l)$ la vraisemblance des données observées y. Étant donné le modèle

M_l, on obtient la loi a posteriori de θ_l en utilisant le théorème de Bayes :

$$\pi(\theta_l|y, M_l) = \frac{\pi(y|\theta_l, M_l)\pi(\theta_l|M_l)}{\int \pi(y|\theta_l, M_l)\pi(\theta_l|M_l)\, d\theta_l}, \tag{2.3}$$

où l'intégrale du dénominateur représente la loi marginale de l'ensemble de données sur toutes les valeurs possibles des paramètres dans le modèle M_l. Cette quantité est appelée la vraisemblance marginale du modèle et est désignée par :

$$\pi(y|M_l) = \int \pi(y|\theta_l, M_l)\pi(\theta_l|M_l)\, d\theta_l. \tag{2.4}$$

Cette intégrale peut en général être difficile à calculer. L'utilisation du critère BIC (Bayesian Information Criterion) comme une approximation à la vraisemblance marginale du modèle peut aider à surmonter ce problème (Clyde 2003) :

$$\pi(y|M_l) \approx \exp\left[\frac{-BIC_{M_l}}{2}\right], \tag{2.5}$$

où : $BIC_{M_l} = -2\log(\pi(y|\theta_l, M_l)) + p_l \log(n)$, avec p_l est la dimension de θ_l.

Soit $\pi(M_l)$ pour $l \in [1, ...,K]$, la loi a priori de chacun des modèles considérés, décrivant l'incertitude préalable sur la capacité de chaque modèle à décrire adéquatement les données. La probabilité a posteriori du modèle M_l, obtenue en utilisant le théorème de Bayes, est comme suit :

$$\pi(M_l|y) = \frac{\pi(y|M_l)\pi(M_l)}{\sum_{k=1}^{K} \pi(y|M_k)\pi(M_k)}. \tag{2.6}$$

Si l'ensemble des modèles considérés $\mathcal{M}$ est discret et fini de taille K, alors on peut assigner une loi a priori non informative qui donne la probabilité $\frac{1}{K}$ à chaque modèle considéré. Ainsi :

$$\pi(M_l) = \pi(M_k) = \frac{1}{K}, \quad \forall(l,k) \in [1, ..., K] \times [1,...,K],$$

dans ce cas la formule (2.6) sera simplifiée à :

$$\pi(M_l|y) = \frac{\pi(y|M_l)}{\sum_{k=1}^{K} \pi(y|M_k)}. \tag{2.7}$$

Si Δ est une quantité d'intérêt présente dans tous les modèles considérés, comme un paramètre, alors sa loi marginale a posteriori est donnée par :

$$\pi(\Delta|y) = \sum_{k=1}^{K} \pi(\Delta|M_k,y)\pi(M_k|y). \tag{2.8}$$

Il s'agit d'une moyenne pondérée de la loi a posteriori de Δ sous chaque modèle, où les coefficients de pondération correspondent aux probabilités a posteriori des modèles.

La moyenne et la variance a posteriori de Δ sont les suivantes :

$$\mathbb{E}[\Delta|y] = \sum_{k=1}^{K} \hat{\Delta}_k \pi(M_k|y), \tag{2.9}$$

$$\mathbb{V}[\Delta|y] = \sum_{k=1}^{K} (\mathbb{V}[\Delta|y,M_k] + \hat{\Delta}_k^2)\pi(M_k|y) - \mathbb{E}[\Delta|y]^2, \tag{2.10}$$

où : $\hat{\Delta}_k = \mathbb{E}[\Delta|y,M_k]$ et $\mathbb{V}[\Delta|y,M_k]$ désignent respectivement l'espérance et la variance a posteriori de Δ sous le modèle M_k.

Chapitre 3

La moyenne bayésienne pour les modèles basés sur les DAG

Les DAG peuvent être utilisés pour identifier les variables confondantes pour lesquelles on peut contrôler pour éliminer la confusion, mais il peut subsister une incertitude quant à la bonne façon de les construire. Dans ce chapitre, on propose une nouvelle méthode de BMA qui se base sur des DAG construits par l'utilisateur en fonction des connaissances du domaine d'application. Étant donné que le BMA traditionnel ne fonctionne pas très bien dans un contexte d'identification de facteurs confondants (Shortreed and Ertefaie 2017, Talbot et al. 2015, Wang et al. 2012), l'hypothèse est que la méthode que nous proposons sera plus performante que le BMA traditionnel puisqu'elle tire mieux profit des connaissances scientifiques a priori en utilisant les DAG comme point de départ.

3.1 La méthode de BMA basé sur les DAG

La méthode de BMA basé sur les DAG est une approche d'analyse de l'ensemble des DAG proposés par un utilisateur à partir du modèle moyen bayésien. L'approche consiste tout d'abord à estimer la probabilité a posteriori de chaque graphe en utili-

sant la vraisemblance marginale des données sous les modèles impliqués par chacun des graphes. Parallèlement, une estimation de l'effet causal pour chaque graphe est déterminée à partir d'approches appropriées en ajustant pour un ensemble optimal de covariables pour éviter le biais de confusion. Finalement, l'effet causal est estimé comme une moyenne pondérée des estimations correspondantes à chacun des graphes, où les poids de pondération sont les probabilités a posteriori de chacun des graphes.

Soit $\mathcal{D} = (D_1, ..., D_K)$, l'ensemble des DAG considérés pour estimer l'effet causal de A sur Y, $L = (L_1, ..., L_Z)$, un ensemble de variables potentiellement confondantes et $O = \{L, A, Y\}$, l'ensemble des données observées. On suppose que tous les DAG considérés ont les mêmes variables.

Les deux éléments fondamentaux sur lesquels s'appuie la méthode de BMA à partir d'un DAG D_k sont : l'effet estimé à partir de ce DAG ($\hat{\Delta}_k$ dans l'équation 2.9) et la probabilité a posteriori du DAG $\pi(D_k|O)$.

Le calcul de l'estimé de l'effet causal de A sur Y pour le DAG D_k, $\hat{\Delta}_k$, nécessite d'abord d'identifier un ensemble d'ajustement suffisant pour contrôler le biais de confusion, par exemple à partir du critère porte-arrière présenté à la sous-section 1.3.4. Dans le but de minimiser la variance, nous proposons l'utilisation d'un ensemble d'ajustement suffisant optimal (Rotnitzky and Smucler 2019, Henckel et al. 2019, Witte et al. 2020). L'identification de cet ensemble peut être réalisée à l'aide de la fonction R `optAdjSet` de la librairie `pcalg` (Perkovic et al. 2017, Witte et al. 2020). Ensuite, il faut estimer l'effet causal à partir d'une méthode statistique en ajustant pour les variables de cet ensemble. Différentes méthodes sont proposées selon le type de l'exposition et de la réponse (voir détails plus bas).

Le calcul de la probabilité a posteriori du DAG D_k, $\pi(D_k|O)$, se fait à partir d'une loi a priori et de la vraisemblance marginale des données. La loi a priori est décidée par

l'utilisateur, mais on suppose une loi a priori uniforme pour simplifier la présentation. Pour calculer la vraisemblance marginale, nous utilisons la propriété markovienne des DAG qui permet une certaine factorisation de la vraisemblance des données. En effet, la densité $f(V)$ de l'ensemble des variables présentes dans le DAG D_k, $V = (V_1, ..., V_{J_k})$, satisfait la factorisation de Markov suivante (Point technique 6.1, Hernán and Robins 2020) :

$$f(V) = \prod_{j=1}^{J_k} f(V_j|P_{jk}),$$

où P_{jk} est l'ensemble des variables parents de la variable V_j dans le DAG D_k et $f(V_j|P_{jk})$ est la densité de la variable V_j en fonction de ses parents.

Notons par V_{jk} le modèle paramétrique ajusté pour la variable V_j présente dans le DAG D_k sur ses variables parents, θ_{jk} les paramètres du modèle V_{jk}, $\pi(V_j|\theta_{jk},P_{jk})$ la vraisemblance des données sous le modèle V_{jk}, M_k le modèle paramétrique des données sous le DAG D_k, θ_k les paramètres du modèle M_k et $\pi(O|\theta_k,M_k)$ la vraisemblance des données sous le modèle M_k.

Remarquons que M_k est simplement la densité jointe impliquée par les modèles paramétriques V_{jk} ajustés pour chacune des variables V_j. Ainsi, on peut écrire :

$$f(V) = \pi(O|\theta_k,M_k) = \prod_{j=1}^{J_k} \pi(V_j|\theta_{jk},P_{jk}).$$

La vraisemblance marginale des données sous le DAG D_k, $\pi(O|D_k)$, qui est déterminée à partir de la vraisemblance des données sous le modèle M_k, $\pi(O|M_k)$, peut s'exprimer comme suit (voir l'équation 2.4) :

$$
\begin{aligned}
\pi(O|D_k) = \pi(O|M_k) &= \int ... \int \prod_{j=1}^{J_k} [\pi(V_j|\theta_{jk},P_{jk})\pi(\theta_{jk})]\, d\theta_{1k}...d\theta_{J_k k} \\
&= \prod_{j=1}^{J_k} \left[\int \pi(V_j|\theta_{jk},P_{jk})\pi(\theta_{jk})\, d\theta_{jk} \right] \\
&= \prod_{j=1}^{J_k} \pi(V_j|P_{jk}) \\
&= \prod_{j=1}^{J_k} \exp\left(\frac{-BIC_{V_{jk}}}{2} \right) \quad \text{(voir l'équation 2.5)} \\
&= \exp\left(\frac{-\sum_{j=1}^{J_k} BIC_{V_{jk}}}{2} \right),
\end{aligned}
$$

tel que $\pi(\theta_k)$, la loi a priori des paramètres θ_k, présume l'indépendance entre les différents paramètres $(\theta_{1k},...,\theta_{J_k k})$, c'est-à-dire $\pi(\theta_k) = \prod_{j=1}^{J_k} \pi(\theta_{jk})$. La factorisation $\pi(O|D_k) = \prod_{j=1}^{J_k} \pi(V_j|P_{jk})$ présente une façon générale de procéder pour calculer la vraisemblance marginale. En effet, l'intégrale pourrait être obtenue de diverses manières (par exemple, par simulation Monte-Carlo, par approximation de Laplace ou de façon analytique...), mais nous avons décidé de faire une approximation BIC dans ce cas. Dans la prochaine section, nous présentons une approche basée sur le BIC pour déterminer la vraisemblance marginale qui ne nécessite pas l'hypothèse d'indépendance des paramètres θ_{jk}.

3.1.1 Calcul de la vraisemblance marginale

Le modèle paramétrique V_{jk}, $\forall j \in [1, ..., J_k]$ et $\forall k \in [1, ..., K]$, peut être par exemple un modèle de régression linéaire de la forme $\mathbb{E}[V_j|P_{jk}, \alpha_{jk}] = \alpha_{0jk} + \alpha_{jk}P_{jk}$ avec une erreur normale de variance constante ou un modèle linéaire généralisé, selon le type de la variable V_j. Dans l'approche que nous proposons, un DAG D_k diffère d'un autre DAG $D_{k'}, \forall (k,k') \in [1, ..., K]$, par les liens considérés dans chacun des deux DAG et non pas par la forme fonctionnelle des modèles ajustés pour les variables V_j présentes dans chacun des DAG. Nous proposons également de calculer la vraisemblance

marginale du modèle M_k, $\pi(O|M_k)$, qui est utilisée pour déterminer la vraisemblance marginale du DAG D_k, $\pi(O|D_k)$, à partir de son BIC. Notons que le BIC de M_k est la somme du BIC de chacun des modèles V_{jk} constituant M_k (Schwarz 1978) :

$$BIC_{M_k} = \sum_{j=1}^{J_k} BIC_{V_{jk}}. \tag{3.1}$$

Démonstration

Soit p_k, la dimension de θ_k et p_{jk}, la dimension de θ_{jk}. Alors :

$$\begin{aligned}
BIC_{M_k} &= -2\log(\pi(O|\theta_k,M_k)) + p_k \log(n) \\
&= -2\log\left(\prod_{j=1}^{J_k} \pi(V_{jk}|\theta_{jk},P_{jk})\right) + p_k \log(n) \\
&= -2\sum_{j=1}^{J_k} \log(\pi(V_{jk}|\theta_{jk},P_{jk})) + p_k \log(n) \\
&= \sum_{j=1}^{J_k} [-2\log(\pi(V_{jk}|\theta_{jk},P_{jk})) + p_{jk} \log(n)] \\
&= \sum_{j=1}^{J_k} BIC_{V_{jk}}.
\end{aligned}$$

Ainsi la vraisemblance marginale du DAG considéré D_k est approximée comme suit :

$$\pi(O|D_k) = \pi(O|M_k) \approx \exp\left[\frac{-BIC_{M_k}}{2}\right]. \tag{3.2}$$

Étant donné que la loi a priori sur les DAG est uniforme, la probabilité a posteriori du DAG D_k est comme suit (voir l'équation 2.7) :

$$\pi(D_k|O) = \frac{\pi(O|D_k)}{\sum_{l=1}^{K} \pi(O|D_l)}.$$

Exemple

Considérons une base de données de taille 100 où les variables sont continues et générées selon une loi normale :

$$L_1 \sim \mathcal{N}(0,1)$$

$$L_2 \sim \mathcal{N}(L_1,1)$$

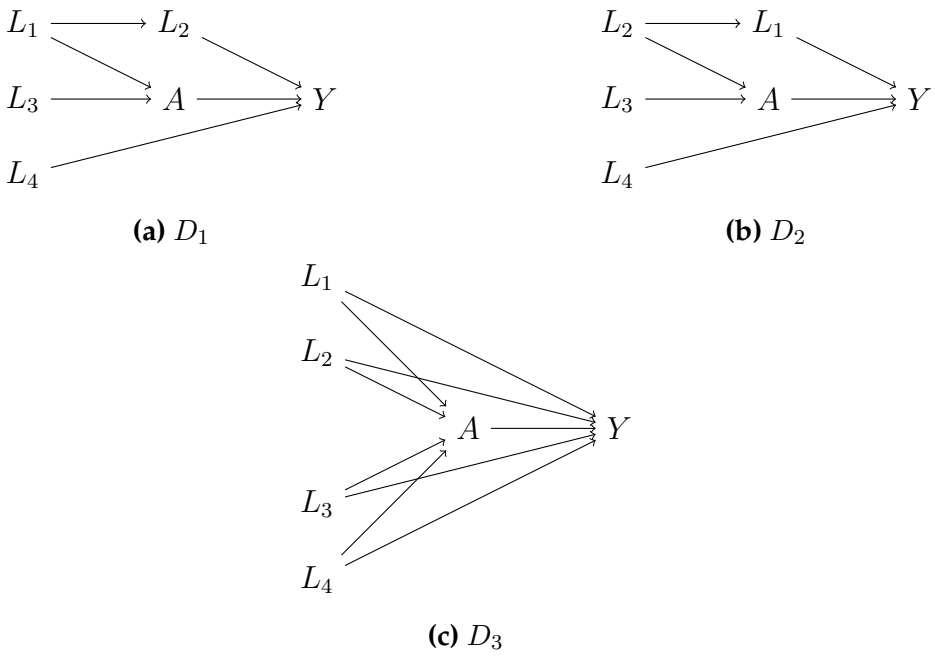
$$L_3 \sim \mathcal{N}(0,1)$$

$$L_4 \sim \mathcal{N}(0,1)$$

$$A \sim \mathcal{N}(L_1 + L_3,1)$$

$$Y \sim \mathcal{N}(A + 0.2L_2 + 0.5L_4,1).$$

Supposons que l'utilisateur donne trois DAG $\mathcal{D} = (D_1, D_2, D_3)$ (figure 3.1) pour estimer l'effet causal de A sur Y, où D_1 est le DAG qui correspond aux équations de génération des données. Soit $L = (L_1, L_2, L_3, L_4)$, un ensemble de variables potentiellement confondantes et $O = \{L, A, Y\}$, l'ensemble des données observées.



(a) D_1 **(b)** D_2

(c) D_3

FIGURE 3.1 – Les graphes proposés pour estimer l'effet causal du traitement A sur la réponse Y. D_1 est le DAG qui correspond aux équations de génération des données.

Appliquons la méthode de BMA basé sur les DAG pour estimer l'effet causal de A sur Y. Pour calculer la probabilité a posteriori de chacun des DAG, nous allons commencer par le calcul de la vraisemblance marginale des données sous chaque DAG.

Les modèles, $V_{jk}\ \forall j \in [1, ..., J_k]$ et $\forall k \in \{1, 2, 3\}$, ajustés pour chaque variable V_j dans chacun des trois DAG sur ses variables parents sont des modèles de régression linéaire.

Les modèles V_{j1} constituant le modèle M_1, lui-même utilisé pour calculer la vraisemblance marginale du DAG D_1, sont :

- $\mathbb{E}[A|P_{11}] = \alpha_{011} + \alpha_{111}L_1 + \alpha_{211}L_3$;
- $\mathbb{E}[L_1|P_{21}] = \alpha_{021}$;
- $\mathbb{E}[L_2|P_{31}] = \alpha_{031} + \alpha_{131}L_1$;
- $\mathbb{E}[L_3|P_{41}] = \alpha_{041}$;
- $\mathbb{E}[L_4|P_{51}] = \alpha_{051}$;
- $\mathbb{E}[Y|P_{61}] = \alpha_{061} + \alpha_{161}A + \alpha_{261}L_2 + \alpha_{361}L_4$.

Les modèles V_{j2} constituant le modèle M_2 sont :

- $\mathbb{E}[A|P_{12}] = \alpha_{012} + \alpha_{112}L_2 + \alpha_{212}L_3$;
- $\mathbb{E}[L_1|P_{22}] = \alpha_{022} + \alpha_{122}L_2$;
- $\mathbb{E}[L_2|P_{32}] = \alpha_{032}$;
- $\mathbb{E}[L_3|P_{42}] = \alpha_{042}$;
- $\mathbb{E}[L_4|P_{52}] = \alpha_{052}$;

- $\mathbb{E}[Y|P_{62}] = \alpha_{062} + \alpha_{162}A + \alpha_{262}L_1 + \alpha_{362}L_4$.

Les modèles V_{j3} constituant le modèle M_3 sont :

- $\mathbb{E}[A|P_{13}] = \alpha_{013} + \alpha_{113}L_1 + \alpha_{213}L_2 + \alpha_{313}L_3 + \alpha_{413}L_4$;
- $\mathbb{E}[L_1|P_{23}] = \alpha_{023}$;
- $\mathbb{E}[L_2|P_{33}] = \alpha_{033}$;
- $\mathbb{E}[L_3|P_{43}] = \alpha_{043}$;
- $\mathbb{E}[L_4|P_{53}] = \alpha_{053}$;
- $\mathbb{E}[Y|P_{63}] = \alpha_{063} + \alpha_{163}A + \alpha_{263}L_1 + \alpha_{363}L_2 + \alpha_{463}L_3 + \alpha_{563}L_4$.

La vraisemblance des données qui sera utilisée pour déterminer le BIC du modèle M_1 (voir l'équation 3.1) est comme suit :

$$\pi(O|\theta_1,M_1) = \prod_{j=1}^{J_1} \pi(V_{j1}|\theta_{j1},P_{j1})$$

où :

$$\pi(V_{11}|\theta_{11},P_{11}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(A_i - \alpha_{011} - \alpha_{111}L_{1i} - \alpha_{211}L_{3i})^2}{2}\right];$$

$$\pi(V_{21}|\theta_{21},P_{21}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(L_{1i} - \alpha_{021})^2}{2}\right];$$

$$\pi(V_{31}|\theta_{31},P_{31}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(L_{2i} - \alpha_{031} - \alpha_{131}L_{1i})^2}{2}\right];$$

$$\pi(V_{41}|\theta_{41},P_{41}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(L_{3i} - \alpha_{041})^2}{2}\right];$$

$$\pi(V_{51}|\theta_{51},P_{51}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(L_{4i} - \alpha_{051})^2}{2}\right];$$

$$\pi(V_{61}|\theta_{61},P_{61}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(Y_i - \alpha_{061} - \alpha_{161}A_i - \alpha_{261}L_{2i} - \alpha_{361}L_{4i})^2}{2}\right].$$

En procédant de la même façon on peut obtenir la vraisemblance de M_2 et M_3. Ainsi, à partir du BIC de ces modèles on peut approximer la vraisemblance marginale $\pi(O|D_k), \forall k \in \{1,2,3\}$, de chacun des trois DAG (voir l'équation 3.2).

Pour calculer la probabilité a posteriori de chacun des trois DAG, il faut préciser une loi a priori. On suppose une loi a priori uniforme :

$$\pi(D_1) = \pi(D_2) = \pi(D_3) = \frac{1}{3}.$$

D'où la probabilité a posteriori $\pi(D_k|O), k \in \{1,2,3\}$, de chacun des trois DAG est comme suit :

- $\pi(D_1|O) = \frac{\pi(O|D_1)}{\sum_{l=1}^{3}\pi(O|D_l)} = 0.60$;
- $\pi(D_2|O) = \frac{\pi(O|D_2)}{\sum_{l=1}^{3}\pi(O|D_l)} = 0.39$;
- $\pi(D_3|O) = \frac{\pi(O|D_3)}{\sum_{l=1}^{3}\pi(O|D_l)} = 4.18 \times 10^{-15}$.

Ensuite, nous allons estimer l'effet causal correspondant à chacun des trois DAG à partir d'une méthode statistique en ajustant pour les variables de l'ensemble d'ajustement optimal (voir la suite de cet exemple).

3.1.2 Estimation de l'effet causal

L'estimation de l'effet causal se fait à partir d'une méthode statistique en ajustant pour les variables de l'ensemble d'ajustement optimal. On propose des méthodes différentes selon le type de l'exposition et de la réponse (le cas où l'exposition est continue et la réponse est binaire n'est pas traité dans le cadre de ce mémoire).

A **continue ou binaire,** Y **continue**

Les modèles ajustés pour estimer l'effet causal correspondant à chacun des DAG considérés, dans le cas d'une exposition continue ou binaire et d'une réponse continue, sont les K modèles de régression linéaire de la forme suivante :

$$\mathbb{E}[Y|A,U_k] = \delta_{0k} + \Delta_k A + \delta_k U_k,$$

où $U_k \subseteq L$, désigne l'ensemble d'ajustement optimal qui permet d'éviter le biais de confusion dans le DAG D_k avec $k \in [1, ..., K]$.

A **binaire,** Y **binaire**

Il est bien connu en inférence causale que la régression logistique implique des rapports de cotes, conditionnel et marginal, non collapsibles (Daniel et al. 2021, Greenland et al. 1999b). En effet, avec une réponse binaire, le rapport de cotes comparant les individus traités et les individus non traités change, même en absence de biais de confusion, après l'inclusion d'une covariable associée avec la réponse dans le modèle.

Pour éviter le problème de collapsibilité, on utilise la méthode de pondération par l'inverse de probabilité de traitement (IPTW) (Chapitres 2 et 12, Hernán and Robins 2020) qui permet d'estimer un rapport de cote marginal ajusté. Les poids spécifiques pour les individus sont définis comme suit :

$$w^A = \frac{1}{P[A|L]}.$$

Pour obtenir une estimation paramétrique de $P[A = 1|L]$, on ajuste un modèle de régression logistique de la variable exposition sur les variables de l'ensemble d'ajustement optimal :

$$\mathbb{E}[A|U_k] = \frac{\exp(\theta_{0k} + \theta_k U_k)}{1 + \exp(\theta_{0k} + \theta_k U_k)}.$$

Les modèles ajustés pour estimer l'effet causal correspondant à chacun des DAG considérés sont les K modèles de régression logistique adaptés par les moindres

carrés pondérés de la forme :

$$\mathbb{E}[Y|A] = \frac{\exp(\delta_{0k} + \Delta_k A)}{1 + \exp(\delta_{0k} + \Delta_k A)}.$$

Estimation de l'effet causal

Au vu de ce qui précède, l'estimation de l'effet causal correspondant au DAG D_k, selon le type de l'exposition et de la réponse, est donc : $\hat{\Delta}_k, \forall k \in [1, ..., K]$.

Ainsi, l'effet causal estimé selon notre approche est :

$$\hat{\Delta} = \sum_{k=1}^{K} \hat{\Delta}_k \pi(D_k|O).$$

Exemple (Suite)

Tout d'abord, nous allons commencer par déterminer un ensemble d'ajustement optimal pour contrôler le biais de confusion pour chacun des trois DAG :

- $U_1 = \{L_2, L_4\}$;
- $U_2 = \{L_1, L_4\}$;
- $U_3 = \{L_1, L_2, L_3, L_4\}$.

Remarquons que U_k représente exactement les mêmes variables que les parents de Y dans le DAG D_k, $\forall k \in \{1, 2, 3\}$. Cependant, ce n'est pas toujours le cas tel qu'illustré dans les scénarios considérés dans la prochaine section.

Étant donné que dans notre exemple les variables exposition et réponse sont continues, alors les modèles ajustés pour estimer l'effet causal sous chacun des trois DAG sont les trois modèles de régression linéaire suivants :

- $\mathbb{E}[Y|A,U_1] = \delta_{01} + \Delta_1 A + \delta_{11} L_2 + \delta_{21} L_4$;
- $\mathbb{E}[Y|A,U_2] = \delta_{02} + \Delta_2 A + \delta_{12} L_1 + \delta_{22} L_4$;

- $\mathbb{E}[Y|A,U_3] = \delta_{03} + \Delta_3 A + \delta_{13} L_1 + \delta_{23} L_2 + \delta_{33} L_3 + \delta_{43} L_4$.

L'estimation de l'effet causal correspondant à chacun des DAG est :

- $\hat{\Delta}_1 = 0.98$;
- $\hat{\Delta}_2 = 0.99$;
- $\hat{\Delta}_3 = 0.99$.

On remarque que les trois estimations sont à peu près égales. Ce résultat s'explique par le fait que les ensembles d'ajustement optimaux U_1, U_2 et U_3 sont des ensembles suffisants pour le vrai DAG D_1.

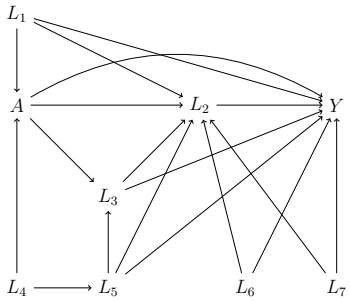
L'effet causal estimé par la méthode de BMA basé sur les DAG pour cet exemple est :

$$\hat{\Delta} = \sum_{k=1}^{3} \hat{\Delta}_k \pi(D_k|O) = 0.99.$$

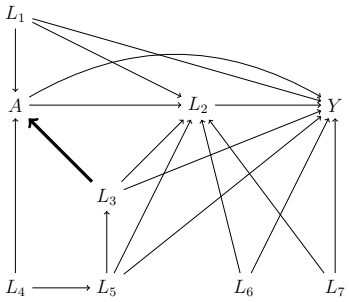
3.2 Étude de simulation

3.2.1 Génération des données

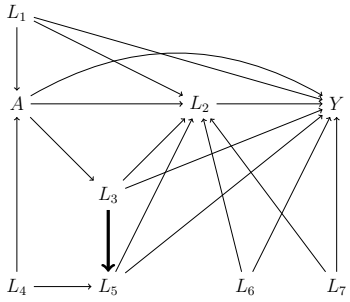
La génération des données de cette étude de simulation est inspirée de l'application qui sera présentée au chapitre 4. L'objectif principal de cette application portera sur l'estimation de l'effet de l'activité physique, A, sur le risque de fractures de la hanche, Y, en utilisant les données de l'étude Study of Osteoporotic Fractures (SOF). Pour créer les DAG de simulation, $\mathcal{D} = (D_1, ..., D_8)$ (figure 3.2), on s'est basé sur D_1 qui représente une structure crédible pour représenter les liens entre les variables disponibles selon nos connaissances du domaine. Les autres DAG ont été obtenus en modifiant le DAG de base selon des incertitudes concernant le rôle de différentes variables. Le mécanisme de génération des données correspond au DAG D_1.



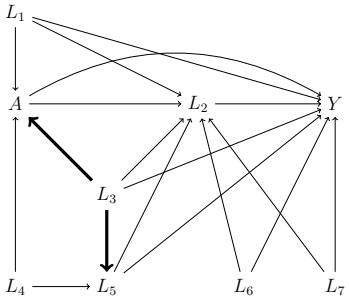
(a) D_1 : DAG de base



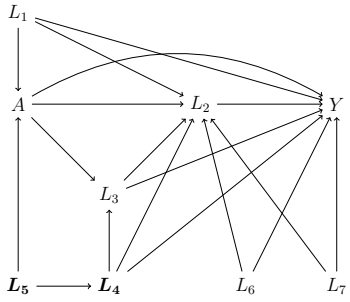
(b) D_2 : L_3 confondant



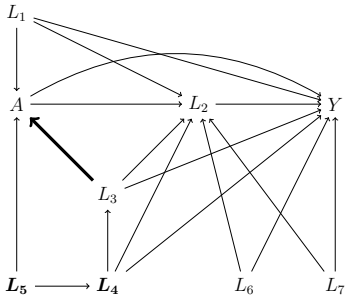
(c) D_3 : L_5 collisionneur



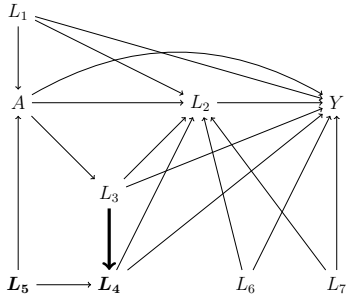
(d) D_4 : L_3 confondant et L_5 collisionneur



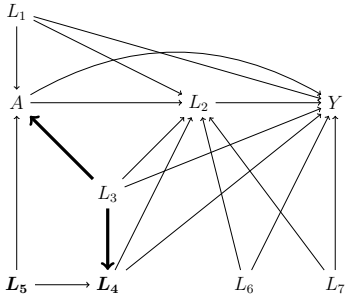
(e) D_5 : L_4-L_5 intervertis



(f) D_6 : L_3 confondant et $L_4 - L_5$ intervertis



(g) D_7 : L_4 collisionneur et $L_4 - L_5$ intervertis



(h) D_8 : L_3 confondant, L_4 collisionneur et $L_4 - L_5$ intervertis

FIGURE 3.2 – Les DAG considérés dans l'étude de simulation basée sur les données générées sous le DAG D_1. Les éléments en gras indiquent les différences par rapport au DAG de base D_1.

On considère une simulation Monte-Carlo impliquant la génération de $n_{sim} = 1000$ jeux de données de taille $n = 1000$. Chaque jeu de données se présente comme $\{O_i = (Y_i, A_i, L_i), \quad i \in [1, ..., n]\}$, pour chacun des 27 mécanismes décrits ci-dessous, où $L_i = (L_{i1}, ..., L_{i7})$. On génère aléatoirement l'ensemble des variables en utilisant une loi normale. Ainsi, pour une base de données $G_h, \forall h \in [1, ..., n_{sim}]$, on a :

$$L_{h1} \sim \mathcal{N}(0,1)$$

$$L_{h4} \sim \mathcal{N}(0,1)$$

$$L_{h6} \sim \mathcal{N}(0,1)$$

$$L_{h7} \sim \mathcal{N}(0,1)$$

$$L_{h5} \sim \mathcal{N}(cL_{h4},1)$$

$$A_h \sim \mathcal{N}(c_A(L_{h1} + L_{h4}),1)$$

$$L_{h3} \sim \mathcal{N}(c(A_h + L_{h5}),1)$$

$$L_{h2} \sim \mathcal{N}(c(A_h + L_{h1} + L_{h3} + L_{h5} + L_{h6} + L_{h7}),1)$$

$$Y_h \sim \mathcal{N}(c_Y(A_h + L_{h1} + L_{h2} + L_{h3} + L_{h5} + L_{h6} + L_{h7}),1).$$

Afin de calculer l'effet total du traitement induit par ce mécanisme de génération de données, on remplace chacune des variables, sauf A, par l'équation qui permet de la générer, jusqu'à ce qu'on obtienne une équation pour Y qui ne dépend que de A et de termes d'erreur :

$$
\begin{aligned}
Y &= c_Y(A + L_1 + L_2 + L_3 + L_5 + L_6 + L_7) + \epsilon_Y \\
&= c_Y(A + \epsilon_{L_1} + c[A + L_1 + L_3 + L_5 + L_6 + L_7] + \epsilon_{L_2} + c[A + L_5] + \epsilon_{L_3} \\
&\qquad + cL_4 + \epsilon_{L_5} + \epsilon_{L_6} + \epsilon_{L_7}) + \epsilon_Y \\
&= c_Y(A + \epsilon_{L_1} + c[A + \epsilon_{L_1} + c[A + c\epsilon_{L_4} + \epsilon_{L_5}] + \epsilon_{L_3} + c\epsilon_{L_4} + \epsilon_{L_5} + \epsilon_{L_6} + \epsilon_{L_7}] \\
&\qquad + \epsilon_{L_2} + c[A + c\epsilon_{L_4} + \epsilon_{L_5}] + \epsilon_{L_3} + c\epsilon_{L_4} + \epsilon_{L_5} + \epsilon_{L_6} + \epsilon_{L_7}) + \epsilon_Y \\
&= c_Y([1 + 2c + c^2]A + c\epsilon_{L_1} + \epsilon_{L_2} + [1 + c]\epsilon_{L_3} + [c + 2c^2 + c^3]\epsilon_{L_4} \\
&\qquad + [1 + 2c + c^2]\epsilon_{L_5} + [1 + c]\epsilon_{L_6} + [1 + c]\epsilon_{L_7}) + \epsilon_Y.
\end{aligned}
$$

Ainsi, pour une augmentation d'une unité en traitement A, le vrai effet du traitement sur la réponse est :

$$\Delta = c_Y(1 + 2c + c^2).$$

Tel que mentionné précédemment, nous considérons 27 mécanismes de génération de données. Ces mécanismes ne diffèrent que dans les valeurs de c, c_A et c_Y. En effet, chacun des trois coefficients peut prendre trois valeurs possibles, soit 1 : un lien fort, 0.4 : un lien moyen et 0.1 : un lien faible.

3.2.2 Analyse des données simulées

Chaque jeu de données simulé est analysé selon la méthode de BMA basé sur les DAG, la méthode de BMA traditionnel et un modèle brut (régression de Y sur A sans l'inclusion des autres variables).

Pour l'application de BMA traditionnel, on a utilisé la fonction `bic.glm()` du package R `BMA` où la probabilité a priori d'inclusion de chacune des variables considérées est 0.5, sauf pour L_2 et L_3. Selon les huit DAG potentiels, L_2 est une variable intermédiaire donc il n'y a pas d'incertitude concernant le rôle de cette variable. Ainsi, on donne une probabilité a priori d'inclusion de 0 à L_2 pour éliminer la possibilité

qu'elle soit sélectionnée. La variable L_3 est confondante dans certains DAG et intermédiaire dans les autres donc il y a de l'incertitude concernant le rôle de cette variable. On considère deux versions différentes de BMA traditionnel afin de refléter l'incertitude par rapport à l'inclusion de cette variable. Dans l'une, la probabilité a priori d'inclusion de L_3 est 0.5 (BMA L_2) et dans l'autre elle est de 0 (BMA L_2, L_3).

Afin d'illustrer comment les estimations peuvent différer en fonction du DAG sélectionné, nous avons, dans un premier temps, appliqué la méthode de BMA basé sur les DAG pour le scénario $c = c_A = c_Y = 1$. Pour ce scénario, nous avons ainsi déterminé, pour chacun des DAG D_k, $\forall k \in [1, ..., 8]$, l'ensemble d'ajustement optimal, l'effet causal estimé, le biais et la probabilité a posteriori pour la méthode de BMA basé sur les DAG.

Dans le but de faciliter l'interprétation des résultats obtenus par ces méthodes d'analyse, il est important de faire une comparaison entre les ensembles d'ajustement suffisants pour le DAG de base D_1 et les ensembles d'ajustement optimaux pour les autres DAG. Parmi les chemins portes-arrières impliqués par le DAG D_1, on trouve $A \leftarrow L_1 \rightarrow Y$, $A \leftarrow L_1 \rightarrow L_2 \rightarrow Y$, $A \leftarrow L_4 \rightarrow L_5 \rightarrow Y$ et $A \leftarrow L_4 \rightarrow L_5 \rightarrow L_3 \rightarrow L_2 \rightarrow Y$. Afin de fermer ces chemins, un ensemble suffisant doit inclure L_1 ainsi que L_4 ou L_5. Par ailleurs, les variables L_2 et L_3 sont intermédiaires. Ces variables ne doivent pas se trouver dans l'ensemble d'ajustement pour qu'il puisse être suffisant. Les variables L_6 et L_7 ne sont pas intermédiaires et elles sont sur des chemins portes-arrières (non détaillés) bloqués par le collisionneur L_2. Par conséquent, leur inclusion dans l'ensemble d'ajustement est sans importance. Les ensembles d'ajustement optimaux pour chacun des DAG sont donnés dans la table 3.1. On remarque que les ensembles pour les DAG D_1, D_3, D_5 et D_7 sont suffisants pour éliminer le biais pour des données générées selon le DAG D_1. Ainsi, bien que les données de notre simulation sont générées selon le DAG D_1, des résultats sans

biais sont attendus si les DAG D_1, D_3, D_5 ou D_7 sont sélectionnés par la procédure de BMA basé sur les DAG.

TABLE 3.1 – Ensembles d'ajustement optimaux des huit DAG considérés.

DAG	Ensemble d'ajustement optimal
D_1	L_1, L_5, L_6 et L_7
D_2	L_1, L_3, L_5, L_6 et L_7
D_3	L_1, L_4, L_6 et L_7
D_4	L_1, L_3, L_5, L_6 et L_7
D_5	L_1, L_4, L_6 et L_7
D_6	L_1, L_3, L_4, L_6 et L_7
D_7	L_1, L_5, L_6 et L_7
D_8	L_1, L_3, L_4, L_6 et L_7

Afin de comparer la performance des méthodes considérées, nous utiliserons les mesures de performance suivantes : le biais, l'écart-type, la racine carrée de l'erreur quadratique moyenne (RMSE), l'erreur-type moyenne et le taux de couverture des intervalles de confiance (crédibilité) à 95 % (voir la table 3.2). Le biais est notre principale mesure de performance d'intérêt.

TABLE 3.2 – Mesures de performance : définitions et estimations.

	Définition	Estimation
Biais	$\mathbb{E}[\hat{\Delta}] - \Delta$	$\frac{\sum_{h=1}^{n_{sim}} \hat{\Delta}_h - \Delta}{n_{sim}}$
Écart-type	$\sqrt{Var(\hat{\Delta})}$	$\sqrt{\frac{\sum_{h=1}^{n_{sim}} (\hat{\Delta}_h - \bar{\Delta})^2}{n_{sim}-1}}$
Erreur-quadratique moyenne	$\mathbb{E}[(\hat{\Delta} - \Delta)^2]$	$\frac{\sum_{h=1}^{n_{sim}} (\hat{\Delta}_h - \Delta)^2}{n_{sim}}$
Erreur-type moyenne	$\sqrt{\mathbb{E}[\widehat{Var}(\hat{\Delta})]}$	$\sqrt{\frac{\sum_{h=1}^{n_{sim}} \widehat{Var}(\hat{\Delta}_h)}{n_{sim}}}$
Couverture	$P(\hat{\Delta}_{inf} \leq \Delta \leq \hat{\Delta}_{sup})$	$\frac{1}{n_{sim}} \sum_{h=1}^{n_{sim}} \mathbb{1}(\hat{\Delta}_{inf,h} \leq \Delta \leq \hat{\Delta}_{sup,h})$

$\bar{\Delta}$ est la moyenne de Δ_h. $\hat{\Delta}_{inf,h} = \hat{\Delta}_h - 1.96 \times \sqrt{Var(\hat{\Delta}_h)}$ et $\hat{\Delta}_{sup,h} = \hat{\Delta}_h + 1.96 \times \sqrt{Var(\hat{\Delta}_h)}$ sont les extrémités des intervalles de confiance de Δ à 95 %.

Pour comprendre les différents résultats, nous allons présenter la probabilité a posteriori d'inclusion de chacune des variables selon les différentes approches considérées

(table 3.5). Les probabilités a posteriori d'inclusion des variables pour le modèle brut sont toutes nulles, sauf A qui a une probabilité a posteriori de 1.

3.2.3 Résultats

Nous présentons dans cette section, les résultats obtenus par les méthodes étudiées pour les différents mécanismes de génération de données. Afin de simplifier la présentation, et étant donné que plusieurs scénarios donnent des résultats qualitativement similaires, seuls les scénarios ayant des résultats différents les uns des autres sont présentés dans les tables de cette section. Nous notons dans le texte les similarités entre les résultats présentés et ceux exclus des tables. Notons par ailleurs que les erreurs-types moyennes sont, en général, similaires à l'écart-type des estimations pour chacune des méthodes dans tous les scénarios. Ces résultats ne sont donc pas rapportés dans les tables présentées dans cette section.

Les résultats de l'analyse préliminaire faite avec la méthode de BMA basé sur les DAG pour le scénario ($c = c_A = c_Y = 1$) sont présentés dans la table 3.3. On remarque que la masse a posteriori est concentrée sur D_1, donc dans ce scénario détaillé le vrai DAG est bien ciblé par notre méthode. Comme signalé auparavant, on constate que le biais obtenu pour les DAG D_1, D_3, D_5 et D_7 est pratiquement nul. En examinant, par exemple, l'ensemble d'ajustement optimal de D_2, on voit qu'il diffère de celui de D_1 seulement dans l'inclusion de L_3. Or, l'effet causal estimé par D_2 est biaisé alors que celui estimé par D_1 est non biaisé. Le biais qu'on observe dans ce cas n'est pas un biais de confusion, mais un biais causé par l'élimination d'une partie de l'effet total. En effet, le vrai effet du traitement dans ce cas fait intervenir non pas seulement l'effet direct de l'exposition sur la réponse, mais aussi l'effet du traitement sur la variable intermédiaire L_3 qui a un impact sur la réponse.

Nous résumons maintenant les résultats complets des simulations pour l'ensemble des 27 scénarios. On peut noter, à partir de la table 3.4, que pour le scénario ($c =$

TABLE 3.3 – Résultats de la simulation avec la méthode de BMA basé sur les DAG dans le scénario des liens forts : $c = c_A = c_Y = 1$.

DAG	Effet causal estimé	Biais	Probabilité a posteriori
D_1	4.00	0.00	1.00
D_2	2.00	-1.99	0.00
D_3	3.99	-0.01	0.00
D_4	2.00	-1.99	0.00
D_5	3.99	-0.01	0.00
D_6	1.00	-2.99	0.00
D_7	4.00	0.00	0.00
D_8	1.00	-2.99	0.00

$c_A = c_Y = 0.1$) toutes les méthodes performent bien. En effet, ces dernières donnent un biais pratiquement nul, des RMSE similaires et des intervalles de confiance à 95 % avec une couverture très proche du niveau attendu (95 %). La même remarque vaut pour les scénarios ($c = c_A = 0.1, c_Y = 0.4$), ($c = 0.1, c_A = 0.4, c_Y = 0.1$), ($c = 0.4, c_A = c_Y = 0.1$).

Pour le scénario ($c = 0.4, c_A = 1, c_Y = 0.1$), le modèle brut est la méthode qui possède le plus de biais. En effet, on remarque un problème de couverture et étant donné que, généralement dans tous les scénarios considérés, l'écart-type correspond à l'erreur-type moyenne pour les différentes méthodes, alors ce problème est causé par le biais. L'approche de BMA basé sur les DAG se démarque avec un biais pratiquement nul, indiquant que cette méthode est la meilleure selon ce critère. La méthode de BMA L_2 a un biais qui peut sembler faible, mais ce biais est suffisant pour induire des problèmes de couverture des intervalles de confiance. On remarque la même chose pour les scénarios ($c = 0.1, c_A = c_Y = 0.4$), ($c = 0.1, c_A = 1, c_Y = 0.4$) et ($c = c_A = 0.4, c_Y = 0.1$).

Tout comme le premier scénario, ($c = c_A = c_Y = 0.1$), les résultats du scénario ($c = 0.1, c_A = 1, c_Y = 0.1$) suggèrent que toutes les méthodes, sauf le modèle brut, ont pu réduire significativement le biais et ont des RMSE pratiquement similaires.

Or, la méthode de BMA basé sur les DAG apparaît comme étant l'approche la plus performante selon le critère de biais et de taux de couverture. Les intervalles de confiance à 95 % produits par cette méthode incluent la vraie valeur 99 % des fois. Quant aux autres approches, les taux de couverture se situent entre 45 % et 90 %. Il convient de souligner que, dans ce scénario, il y a une une certaine différence entre l'écart-type et l'erreur-type moyenne pour toutes les méthodes, sauf le modèle brut. Plus précisément, il y a une sous-estimation de la variance de l'estimateur pour la méthode de BMA traditionnel et une surestimation pour le BMA basé sur les DAG.

Pour le scénario ($c = 0.4, c_A = 0.1, c_Y = 1$), les méthodes BMA DAG et BMA L_2, L_3 performent bien selon le critère de biais et de RMSE. En effet, ces deux approches ont un biais très proche de zéro, des RMSE faibles et de bons taux de couverture. Il convient de souligner que la méthode de BMA L_2 et le modèle brut n'ont pas réussi à éliminer le biais et en raison de ce biais, les taux de couverture résultant de ces deux approches sont très mauvais (entre 0 % et 53 %). On peut dire la même chose pour les scénarios ($c = c_A = 0.1, c_Y = 1$), ($c = 0.4, c_A = 0.1, c_Y = 0.4$), ($c = 1, c_A = c_Y = 0.1$), ($c = 1, c_A = 0.1, c_Y = 0.4$), ($c = 1, c_A = 0.1, c_Y = 1$).

Pour le scénario ($c = c_A = c_Y = 0.4$), les méthodes BMA DAG et BMA L_2, L_3 sont celles qui réduisent le plus le biais et le RMSE. Les taux de couverture résultant de l'utilisation du modèle brut et du BMA L_2 sont très mauvais (pratiquement 0 %). Quant aux autres approches, les taux de couverture sont à 95 %. La même remarque vaut pour les scénarios ($c = 0.1, c_A = c_Y = 1$), ($c = 0.1, c_A = 0.4, c_Y = 1$), ($c = 0.4, c_A = 1, c_Y = 0.4$), ($c = 0.4, c_A = c_Y = 1$), ($c = c_A = 0.4, c_Y = 1$), ($c = 1, c_A = 0.4, c_Y = 0.1$), ($c = 1, c_A = c_Y = 0.4$), ($c = 1, c_A = 0.4, c_Y = 1$), ($c = c_A = 1, c_Y = 0.1$), ($c = c_A = 1, c_Y = 0.4$) et ($c = c_A = c_Y = 1$).

En résumé, les résultats de simulation suggèrent que, en général, lorsque les coefficients c_Y et c_A sont faibles à moyens (entre 0.1 et 0.4) toutes les méthodes considérées fonctionnent bien. Pour les scénarios où seul le modèle brut ne fonctionne pas bien, le

coefficient c_Y prend généralement des valeurs faibles à moyennes et le coefficient c_A prend des valeurs moyennes à fortes. Pour les scénarios où le BMA L_2 ne fonctionne pas bien c_Y prend des valeurs moyennes à fortes et le coefficient c_A prend une valeur faible. Finalement, en général, avec des liens qui ont tendance à être forts (c_Y et c_A entre 0.4 et 1) le BMA DAG et le BMA L_2, L_3 fonctionnent bien.

Il convient de souligner que la méthode de BMA L_2, L_3 performe bien dans l'ensemble de ces scénarios, car nous avons donné une probabilité a priori d'inclusion de 0 à la variable L_3 (voir la sous-section 3.2.2). Or, le BMA traditionnel (la version BMA L_2 dans ce cas) performe généralement moins bien que le BMA basé sur les DAG puisqu'on se trompe sur le rôle de la variable L_3 et on donne une probabilité a priori d'inclusion non nulle à cette variable.

TABLE 3.4 – Résultats des simulations obtenues avec différents mécanismes pour les méthodes considérées.

Mécanisme	Méthode	Vrai effet	Biais	Écart-type	RMSE	Couverture
$c = 0.1$	Brut	0.121	0.01	0.03	0.03	0.93
$c_A = 0.1$	BMA DAG		0.00	0.03	0.03	0.95
$c_Y = 0.1$	BMA L_2		-0.01	0.03	0.03	0.94
	BMA L_2, L_3		0.00	0.03	0.03	0.95
$c = 0.4$	Brut	0.196	0.07	0.02	0.08	0.03
$c_A = 1$	BMA DAG		0.00	0.02	0.02	0.96
$c_Y = 0.1$	BMA L_2		-0.04	0.03	0.05	0.66
	BMA L_2, L_3		0.01	0.03	0.03	0.91
$c = 0.1$	Brut	0.121	0.04	0.02	0.04	0.45
$c_A = 1$	BMA DAG		0.00	0.03	0.03	0.99
$c_Y = 0.1$	BMA L_2		0.00	0.03	0.03	0.90
	BMA L_2, L_3		0.01	0.03	0.03	0.85
$c = 0.4$	Brut	1.960	0.22	0.12	0.25	0.53
$c_A = 0.1$	BMA DAG		0.00	0.07	0.07	0.97
$c_Y = 1$	BMA L_2		-0.56	0.05	0.56	0.00
	BMA L_2, L_3		0.00	0.06	0.06	0.96
$c = 0.4$	Brut	0.784	0.26	0.05	0.27	0.00
$c_A = 0.4$	BMA DAG		0.00	0.04	0.04	0.95
$c_Y = 0.4$	BMA L_2		-0.22	0.04	0.23	0.00
	BMA L_2, L_3		0.00	0.04	0.04	0.95

TABLE 3.5 – Les probabilités a posteriori des variables selon les trois approches considérées.

Mécanisme	Méthode	Probabilité a posteriori d'inclusion							
		A	L_1	L_2	L_3	L_4	L_5	L_6	L_7
$c = 0.1$	BMA DAG	1.00	1.00	0.00	0.45	0.25	0.75	1.00	1.00
$c_A = 0.1$	BMA L_2	1.00	0.76	0.00	0.80	0.03	0.78	0.78	0.78
$c_Y = 0.1$	BMA L_2, L_3	1.00	0.76	0.00	0.00	0.03	0.85	0.78	0.77
$c = 0.4$	BMA DAG	1.00	1.00	0.00	0.00	0.00	0.99	1.00	1.00
$c_A = 1$	BMA L_2	1.00	0.76	0.00	0.94	0.05	0.91	0.95	0.93
$c_Y = 0.1$	BMA L_2, L_3	1.00	0.76	0.00	0.00	0.06	0.99	0.95	0.93
$c = 0.1$	BMA DAG	1.00	1.00	0.00	0.01	0.46	0.54	1.00	1.00
$c_A = 1$	BMA L_2	1.00	0.56	0.00	0.77	0.06	0.77	0.77	0.78
$c_Y = 0.1$	BMA L_2, L_3	1.00	0.56	0.00	0.00	0.06	0.84	0.76	0.78
$c = 0.4$	BMA DAG	1.00	1.00	0.00	0.01	0.00	0.99	1.00	1.00
$c_A = 0.1$	BMA L_2	1.00	1.00	0.00	1.00	0.04	1.00	1.00	1.00
$c_Y = 1$	BMA L_2, L_3	1.00	1.00	0.00	0.00	0.04	1.00	1.00	1.00
$c = 0.4$	BMA DAG	1.00	1.00	0.00	0.00	0.00	0.99	1.00	1.00
$c_A = 0.4$	BMA L_2	1.00	1.00	0.00	1.00	0.04	1.00	1.00	1.00
$c_Y = 0.4$	BMA L_2, L_3	1.00	1.00	0.00	0.00	0.04	1.00	1.00	1.00

En observant la table 3.5, on constate que la méthode de BMA basé sur les DAG permet de distinguer une variable intermédiaire et un confondant dans cet exemple. En effet, la probabilité a posteriori donnée à L_2 est toujours nulle et la probabilité a posteriori donnée à L_3 est relativement faible (inférieure à 0.5). Or, le BMA L_2 donne toujours une probabilité a posteriori élevée à L_3 (entre 0.77 et 1). Par ailleurs, L_1 est toujours incluse dans le BMA DAG par construction, un ensemble suffisant doit toujours inclure L_1, mais pas dans les autres méthodes. Les variables L_4 ou L_5 sont aussi toujours incluses dans le BMA DAG, car un ensemble suffisant doit contenir l'une ou l'autre de ces deux variables pour bloquer les chemins portes-arrières impliqués par le DAG D_1. La seule source de biais possible pour cette méthode est donc l'inclusion de L_2 ou L_3.

Analyse de sensibilité

Afin de tester la robustesse des approches considérées en présence d'une mauvaise spécification du modèle de réponse, nous avons repris les simulations avec des modifications dans la formule de génération de la variable réponse. Le premier cas considéré est $Y_h \sim \mathcal{N}(c_Y(A_h + L_{h1} + L_{h2} + L_{h3} + L_{h5} + L_{h6} + L_{h7}) + c_i L_{h4} L_{h5} + c_{ii} L_{h1}^2, 1)$ et le second cas est $Y_h \sim \mathcal{N}(c_Y(A_h + L_{h1} + L_{h2} + L_{h3} + L_{h5} + L_{h6} + L_{h7}) + c_i L_{h6} L_{h7} + c_{ii} L_{h1}^2 + c_{ii} L_{h4}^2 + c_{ii} L_{h5}^2, 1)$. Il convient de noter que les modèles utilisés dans les méthodes BMA n'incluent pas de termes d'interaction ou quadratiques mais seulement des formes fonctionnelles linéaires.

À partir des résultats présentés dans la table 3.6, la conclusion par rapport aux approches considérées reste inchangée. En effet, même si la spécification du modèle n'est pas correcte par rapport à certaines variables confondantes, qui ne sont pas liées à la variable exposition, les résultats obtenus correspondent pratiquement aux résultats obtenus avant d'apporter ces modifications.

TABLE 3.6 – Résultats des simulations obtenues en présence d'une mauvaise spécification du modèle de réponse.

Cas	Mécanisme	Méthode	Vrai effet	Biais	Écart-type	RMSE	Couverture
1er cas	$c = 0.4$	Brut	1.960	0.21	0.12	0.27	0.60
	$c_A = 0.1$	BMA DAG		0.00	0.08	0.08	0.95
	$c_Y = 1$	BMA L_2		-0.56	0.05	0.56	0.00
	$c_i = c_{ii} = 0.1$	BMA L_2, L_3		0.00	0.06	0.06	0.94
2ème cas	$c = 0.4$	Brut	1.960	0.22	0.12	0.25	0.54
	$c_A = 0.1$	BMA DAG		-0.01	0.07	0.07	0.97
	$c_Y = 1$	BMA L_2		-0.58	0.05	0.56	0.00
	$c_i = c_{ii} = 0.1$	BMA L_2, L_3		0.00	0.06	0.06	0.96
1er cas	$c = 0.4$	Brut	0.784	0.27	0.05	0.27	0.00
	$c_A = 0.4$	BMA DAG		0.00	0.04	0.04	0.96
	$c_Y = 0.4$	BMA L_2		-0.22	0.03	0.23	0.00
	$c_i = c_{ii} = 0.1$	BMA L_2, L_3		0.00	0.04	0.04	0.96
2ème cas	$c = 0.4$	Brut	0.784	0.26	0.05	0.27	0.00
	$c_A = 0.4$	BMA DAG		0.00	0.04	0.04	0.96
	$c_Y = 0.4$	BMA L_2		-0.22	0.03	0.23	0.00
	$c_i = c_{ii} = 0.1$	BMA L_2, L_3		0.00	0.04	0.04	0.96

Chapitre 4

Application

Dans ce chapitre nous utilisons les données de l'étude Study of Osteoporotic Fractures (SOF) pour estimer l'effet causal de la pratique de l'activité physique sur le risque de fractures de la hanche. L'étude SOF est une étude observationnelle d'une cohorte de 10 366 femmes âgées de 65 ans ou plus qui ont été recrutées à partir de quatre régions des États-Unis : Baltimore, Pittsburgh, Minneapolis et Portland. En 1986, 9704 femmes caucasiennes ont été engagées dans le cadre de cette étude. En 1997, 662 femmes afro-américaines ont été ajoutées à la cohorte. Puisque les facteurs de risque des fractures diffèrent selon la race, l'échantillon utilisé dans ce mémoire contient seulement les femmes caucasiennes n'ayant pas reçu de diagnostic d'ostéoporose de leur médecin avant l'examen au recrutement en 1986 pour simplifier la présentation. Ainsi, seules les informations récoltées au recrutement en 1986, ainsi que le suivi sur 10 ans des fractures de la hanche sont considérées pour ce projet. Notons qu'une exemption d'approbation éthique a été obtenue du Comité d'éthique de la recherche du CHU de Québec - Université Laval (dossier #2021-5786).

4.1 Description des données

4.1.1 La variable réponse

La variable réponse est obtenue en vérifiant si la participante a eu une fracture de la hanche dans les 10 années suivant sa visite de recrutement (0 = non, 1 = oui). La mesure est rapportée par la participante en fonction d'un suivi effectué à tous les 4 mois et vérifiée en clinique par un médecin si la participante rapporte une fracture.

4.1.2 La variable exposition

La variable exposition est la pratique d'une activité physique dans la semaine précédente (0 = non, 1 = oui). En effet, on vérifie si la participante a pratiqué, au moins une fois par semaine, une activité physique régulière (marche rapide, jogging, vélo, etc.) suffisamment longue pour transpirer.

4.1.3 Les variables confondantes

Les variables potentiellement confondantes considérées dans cette application incluent l'âge, le poids (à 25 ans, à 50 ans et au recrutement), la taille, l'indice de masse corporelle IMC (calculé en fonction de la taille et du poids de la participante au recrutement), le calcium (la quantité moyenne de calcium consommé dans une semaine), l'éducation (le nombre d'années d'études effectuées), la consommation d'alcool (le nombre de consommations prises par semaine multiplié par le nombre d'années depuis lesquelles la participante boit), le tabagisme (la quantification du tabagisme est calculée comme le produit du nombre de paquets de cigarettes fumés par jour par le nombre d'années depuis lesquelles la participante est fumeuse), le statut de fumeur (0 = jamais, 1 = passé ou actuel), les médicaments stéroïdiens (0 = jamais, 1 = passé ou actuel) et les variables d'antécédents familiaux : est-ce que la mère a déjà eu

une fracture, peu importe laquelle, (0 = non, 1 = oui), est-ce que le père a déjà eu une fracture, peu importe laquelle, (0 = non, 1 = oui).

4.1.4 Les variables intermédiaires

La densité de l'extrémité proximale de l'os radius (PRXBMD) est mesurée par absorption photonique (en mg/cm^2) et elle est la seule variable intermédiaire considérée dans cette application.

4.1.5 Statistiques descriptives

Notre base de données comporte des valeurs manquantes sur de nombreuses variables. La variable réponse comporte le plus grand nombre de données manquantes (20.74 %) alors que la variable exposition contient seulement 0.07 % de données manquantes. Les autres variables ont des pourcentages de valeurs manquantes entre 0 % (les variables d'antécédents familiaux) et 2.65 % (le poids à 50 ans). Le pourcentage des observations qui ont au moins une valeur manquante (1976 observations) est de 25.43 %. Il existe globalement 2394 valeurs manquantes, soit 1.93 % des données.

La table 4.1 présente les statistiques descriptives de l'ensemble des variables de notre base de données. Les moyennes et les écarts-types sont rapportés pour les variables continues. Les fréquences et les pourcentages d'observations sont rapportés pour les variables catégorielles. Le regroupement est fait par rapport aux deux catégories d'exposition. Les observations pour lesquelles la variable d'exposition est manquante (n = 6) ont été préalablement éliminées pour la construction de cette table.

Notons que seulement 7.3 % des participantes ont une fracture. La réponse est rare donc les rapports de cotes peuvent être interprétés comme un risque relatif (voir 4.3).

TABLE 4.1 – Les statistiques descriptives obtenues selon le type des variables étudiées en fonction des niveaux de la variable traitement activité physique.

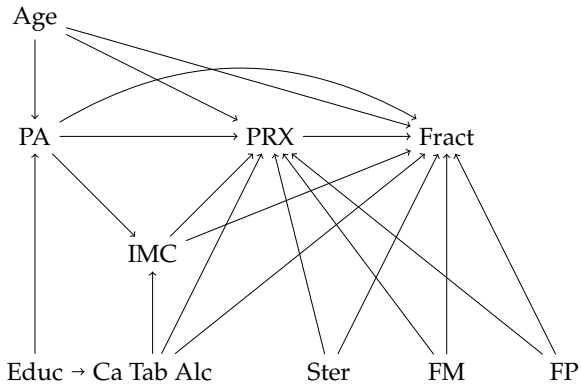
	Activité physique	
	Non	Oui
n	2425	5340
Âge (années)	72.43 (5.59)	70.99 (4.86)
Poids à 25 ans (kg)	56.35 (7.17)	56.22 (6.55)
Poids à 50 ans (kg)	63.44 (8.91)	61.95 (8.03)
Poids (kg)	68.91 (12.98)	66.64 (11.41)
Taille (cm)	158.66 (5.81)	159.85 (5.73)
IMC (kg/m^2)	27.36 (4.83)	26.07 (4.19)
Calcium (mg)	4609.60 (2788.71)	5067.71 (2914.45)
Éducation (années)	11.76 (2.65)	12.95 (2.71)
Alcool	89.61 (216.73)	99.37 (214.07)
Tabagisme	12.20 (23.06)	9.47 (17.98)
Statut de fumeur		
Jamais	1481 (61.20)	3250 (61.10)
Passé ou actuel	938 (38.80)	2070 (38.90)
Stéroïde		
Jamais	39 (1.60)	82 (1.60)
Passé ou actuel	2343 (98.40)	5154 (98.40)
Fracture de la mère		
Non	1772 (73.10)	3855 (72.20)
Oui	653 (26.90)	1485 (27.80)
Fracture du père		
Non	2083 (85.90)	4531 (84.90)
Oui	342 (14.10)	809 (15.10)
PRXBMD (mg/cm^2)	0.63 (0.11)	0.64 (0.10)

⋆ On présente les fréquences et les pourcentages d'observations pour les variables binaires, la moyenne et l'écart-type pour les variables continues.

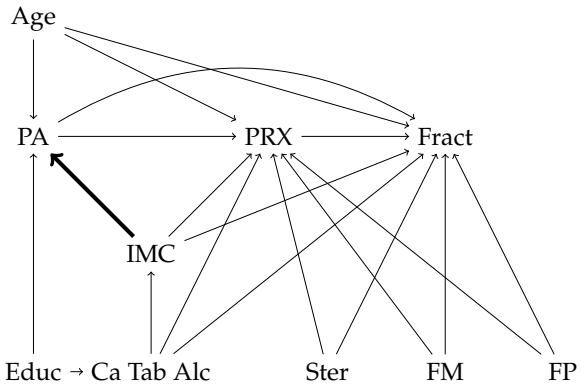
L'ensemble des DAG considérés dans cette application est présenté dans la figure 4.1. Le DAG de base représente une structure crédible pour décrire les relations présumées entre les variables disponibles selon nos connaissances du domaine. Les deux autres DAG ont été obtenus en modifiant le DAG de base selon des incertitudes concernant le rôle de la variable IMC. Cette dernière peut être (a) une variable intermédiaire qui est affectée par la pratique d'une activité physique et qui peut avoir un effet sur la densité osseuse PRXBMD, (b) un confondant qui affecte la pratique d'une activité

physique et la densité osseuse PRXBMD ou (c) un effet commun de la pratique d'une activité physique et de PRXBMD.

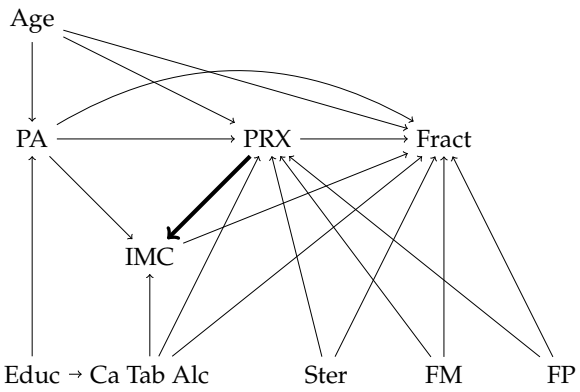
Pour alléger la présentation, les flèches vers les trois variables calcium, tabagisme et alcool et celles à partir de ces variables ne sont présentées qu'une seule fois dans la figure 4.1. Les abréviations utilisées dans cette figure sont PA : La pratique d'une activité physique, PRX : PRXBMD, Fract : Fracture de la hanche, Educ : Éducation, Ca Tab Alc : Calcium Tabagisme Alcool, Ster : Stéroïde, FM : Fracture de la mère et FP : Fracture du père.



(a) Le DAG de base



(b) IMC est confondant



(c) IMC est collisionneur

FIGURE 4.1 – Les DAG considérés dans l'étude de l'effet causal de la pratique de l'activité physique (PA) sur le risque de fractures de la hanche (Fract). Les éléments en gras indiquent les différences par rapport au DAG de base.

4.2 Pré-traitement des données

4.2.1 Valeurs aberrantes

Les valeurs aberrantes sont des observations qui semblent incompatibles avec le reste de l'ensemble des données (Barnett and Lewis 1994). Ces valeurs résultent de divers facteurs, dont les erreurs de réponse des participants et les erreurs de saisie des données. Pour détecter les valeurs aberrantes dans les données, nous avons utilisé la fonction `rosnerTest()` du package R `EnvStats`. Cette fonction permet d'appliquer le test de Rosner (Rosner 1975, Rosner 1983) pour identifier les valeurs aberrantes dans l'ensemble de données considéré. L'approche de Rosner est conçue pour éviter le problème où une valeur aberrante proche d'une autre valeur aberrante passe inaperçue. Le test de Rosner n'est approprié que lorsque les données, à l'exclusion des valeurs aberrantes suspectées, proviennent d'une distribution normale. Les données ne doivent pas être exclues de l'analyse uniquement sur la base des résultats de ce test ou de tout autre test statistique. Si des valeurs sont signalées comme pouvant être aberrantes, il est recommandé de procéder à une enquête plus approfondie afin de déterminer s'il existe une explication plausible qui justifie leur suppression ou leur remplacement. Nous avons d'abord appliqué différentes transformations aux variables continues qui contiennent des valeurs aberrantes suspectées afin de les normaliser. Nous avons appliqué le logarithme pour les variables : alcool, PRXBMD, tabagisme et poids à 25 ans et la racine carrée pour les variables : poids à 50 ans, poids au recrutement, taille, IMC, éducation et calcium. Ensuite, nous avons appliqué le test de Rosner sur les variables transformées et les valeurs aberrantes détectées sont deux observations de la variable calcium. Par souci de prudence, nous avons comparé ces deux observations avec le $3 \times 99^{\text{ème}}$ centile de la variable calcium (Kwak and Kim 2017, Liao et al. 2016) et comme les deux observations sont inférieures à cette valeur, nous avons jugé qu'il serait utile de les garder dans notre jeu de données.

4.2.2 Valeurs manquantes

Afin de contourner le problème des valeurs manquantes dans les données, nous avons utilisé l'imputation multiple (Rubin 1996, Allison 2000, Rubin 2004). En effet, l'imputation multiple permet de prendre en compte l'incertitude de prédiction des données manquantes en imputant `b` valeurs plausibles pour chaque valeur non observée dans les données. On obtient ainsi `b` versions différentes de bases de données, où les données non manquantes sont identiques, mais où les entrées de données manquantes diffèrent. Il est généralement recommandé d'utiliser le pourcentage moyen de données manquantes dans le jeu de données comme nombre d'imputations (Bodner 2008, White et al. 2011). Ainsi, comme il y a 25 % de données manquantes en moyenne dans notre jeu de données alors nous avons considéré `b` = 25 bases de données imputées. Pour faire l'imputation, nous avons utilisé la fonction `mice()` du package R `mice`. La fonction `mice()` prend en compte tous les aspects de l'incertitude relative aux données manquantes en comparaison aux méthodes d'imputation simple (l'imputation par la moyenne, la médiane ou le mode). Cette fonction suppose que les données manquantes sont de type MAR (Missing At Random), ce qui signifie que la probabilité qu'une valeur soit manquante ne dépend que des valeurs observées sur certaines covariables et peut être prédite à partir de celles-ci. Pour imputer les valeurs manquantes, le package `mice` utilise un algorithme de manière à utiliser les informations des autres variables de l'ensemble de données pour prédire et imputer les valeurs manquantes. Il existe des méthodes spécifiques à chaque type de variable. Dans notre cas, l'appariement prédictif des moyennes (Rubin 1986, Little 1988) est utilisé pour prédire les valeurs manquantes continues et la régression logistique est utilisée pour les valeurs manquantes binaires.

4.3 Application

Les données sont analysées à l'aide des méthodes considérées dans le chapitre 3 : BMA DAG, BMA traditionnel et le modèle brut. Il convient de signaler que nous avons utilisé les variables transformées décrites à la sous-section 4.2.1 dans ces analyses.

Pour la méthode de BMA DAG, et étant donné que l'exposition et la réponse sont des variables binaires, nous avons utilisé la méthode de pondération par l'inverse de probabilité de traitement (*Inverse Probability of Treatment Weighting* - IPTW) pour éviter le problème de collapsibilité. Ensuite pour estimer l'effet causal pour chacun des DAG considérés, nous avons ajusté des modèles de régression logistique adaptés par les moindres carrés pondérés (voir la sous-section 3.1.2).

Pour le BMA traditionnel, nous avons utilisé la fonction `bic.glm()` du package R `BMA` avec le lien logit. La réponse est ajustée sur l'exposition avec une probabilité a priori d'inclusion de 1 et les variables potentiellement confondantes avec une probabilité a priori d'inclusion de 0.5, sauf pour PRXBMD et IMC. Selon les trois DAG considérés, PRXBMD est une variable intermédiaire donc il n'y a pas d'incertitude concernant le rôle de cette variable. Par conséquent, une probabilité a priori d'inclusion de 0 est donnée à PRXBMD afin d'éliminer la possibilité qu'elle soit sélectionnée. La variable IMC est confondante dans certains DAG et intermédiaire dans les autres donc il y a de l'incertitude concernant le rôle de cette variable. Ainsi, on considère deux versions différentes de BMA traditionnel afin de refléter l'incertitude par rapport à l'inclusion de cette variable. Dans une première version, notée par BMA PRX, la probabilité a priori d'inclusion de IMC est 0.5. Dans une deuxième version, notée par BMA PRX IMC, la probabilité a priori d'inclusion de IMC est 0.

Pour le modèle brut, nous avons ajusté un modèle logistique où la réponse, le risque de fractures de la hanche, est ajustée sur l'exposition, la pratique d'une activité physique, sans l'inclusion des autres variables.

Nous avons appliqué les méthodes considérées sur les 25 ensembles de données obtenus séparément et nous avons combiné leurs résultats (Rubin 1996). L'effet causal estimé pour chaque méthode est calculé comme la moyenne des 25 estimations obtenues, $\hat{\Delta}_1, ..., \hat{\Delta}_b$, avec les variances estimées correspondantes $\hat{v}_1, ..., \hat{v}_b$:

$$\hat{\Delta} = \frac{1}{b} \sum_{m=1}^{b} \hat{\Delta}_m.$$

L'estimation de la variance totale de l'effet causal estimé pour chaque méthode est calculée comme suit :

$$\hat{T} = \hat{w} + \left(1 + \frac{1}{b}\right) \hat{z},$$

où : $\hat{w} = \frac{1}{b} \sum_{m=1}^{b} \hat{v}_m$, est la variance intra-imputation et $\hat{z} = \frac{1}{b-1} \sum_{m=1}^{b} (\hat{\Delta}_m - \hat{\Delta})^2$, est la variance inter-imputation qui mesure l'incertitude due à l'imputation.

L'intervalle de confiance à 95 % de l'effet causal, Δ, pour chaque méthode est de la forme suivante :

$$IC = \hat{\Delta} \pm 1.96 \times \sqrt{\hat{T}}.$$

La table 4.2 résume les rapports de cotes estimés correspondants à l'effet causal de la pratique d'une activité physique sur le risque de fractures de la hanche, leurs intervalles de confiance à 95 % ainsi que les écarts-types des estimations de l'effet causal de l'exposition sur la réponse pour les méthodes ci-dessus.

Le rapport de cotes pour le modèle brut est 0.59 avec un intervalle de confiance de [0.48 , 0.71]. Ce résultat indique que les chances d'avoir une fracture de la hanche sont 41 % moins élevées pour les participantes qui pratiquent une activité physique que pour celles qui ne pratiquent aucune activité physique.

Pour la méthode de BMA DAG, la masse a posteriori est concentrée sur le DAG de base, qui diffère des deux autres DAG en considérant l'IMC comme une variable intermédiaire plutôt qu'un confondant ou un effet commun. La probabilité a posteriori de ce DAG est 0.96. La méthode de BMA DAG estime que pour les femmes qui

pratiquent une activité physique, le risque d'avoir une fracture de la hanche est inférieur de 21 % (le rapport de cotes est 0.79 avec un intervalle de confiance de [0.65 , 0.97]) par rapport à celles qui ne pratiquent aucune activité physique, toutes les autres caractéristiques restant inchangées.

Les méthodes de BMA traditionnel BMA PRX et BMA PRX IMC estiment que le risque d'avoir une fracture de la hanche est respectivement de 26 % (le rapport de cotes est 0.74 avec un intervalle de confiance de [0.59 , 0.93]) et 25 % (le rapport de cotes est 0.75 avec un intervalle de confiance de [0.60 , 0.94]) moins élevé lorsque la participante pratique une activité physique.

Le BMA traditionnel est sujet à différentes limites dont le problème de collapsibilité qui n'est pas géré adéquatement dans la fonction `bic.glm()`. Par conséquent, la comparaison entre cette méthode et la méthode de BMA DAG est difficile à effectuer car le rapport de cotes est marginal pour une méthode alors qu'il est conditionnel pour l'autre. Ainsi, les différences entre les résultats de ces méthodes sont dues à la fois aux probabilités attribuées qui peuvent être différentes, mais aussi au fait que les estimations ne sont pas comparables. Bien que les résultats de ces méthodes sont difficiles à comparer, il faut noter que la valeur ajoutée du BMA DAG réside dans l'information à propos du rôle de la variable IMC dans cette application.

TABLE 4.2 – Les rapports de cotes estimés pour l'effet causal de la pratique d'une activité physique sur le risque de fractures de la hanche, ainsi que les intervalles de confiance à 95 % et les écarts-types correspondants.

Méthode	Rapport de cotes	Écart-type	Intervalle de confiance à 95 %	
			Borne inférieure	Borne supérieure
Brut	0.59	0.10	0.48	0.71
BMA DAG	0.79	0.10	0.65	0.97
BMA PRX	0.74	0.12	0.59	0.93
BMA PRX IMC	0.75	0.12	0.60	0.94

Conclusion

Bien que le modèle moyen bayésien traditionnel est efficace dans certains cas, les implémentations de cette méthode peuvent être confrontées à certaines limitations dans un contexte de sélection des facteurs de confusion. Dans ce mémoire, nous nous sommes intéressés à développer une nouvelle approche d'analyse utilisant l'ensemble des graphes acycliques orientés proposés par un utilisateur à partir du modèle moyen bayésien. En effet, cette méthode tire mieux profit des connaissances antérieures et du processus de génération de données en utilisant les DAG comme point de départ.

Dans un premier temps, nous avons commencé par examiner les concepts et les outils utilisés en inférence causale ainsi qu'en statistique bayésienne. Ensuite, nous avons présenté de façon détaillée notre approche, ainsi qu'une étude de simulation de type Monte-Carlo. Finalement, nous avons présenté les résultats obtenus à partir de l'application de notre approche sur les données de l'étude Study of Osteoporotic Fractures (SOF).

Les résultats de l'étude de simulation vont dans le sens de valider l'hypothèse que la nouvelle approche que nous avons développée est plus performante que le BMA traditionnel dans certains scénarios. En comparaison au BMA traditionnel, le BMA DAG a globalement réussi à réduire davantage le biais et l'erreur quadratique moyenne, particulièrement quand le rôle d'une variable était incertain et qu'on attribuait une probabilité a priori d'inclusion non nulle à une variable qui aurait de facto due être exclue. De plus, les résultats obtenus dans le chapitre 4 nous permettent de voir le

bénéfice potentiel de cette méthode. En effet, tel que vu dans ce chapitre, la méthode de BMA DAG permet d'avoir de l'information concernant le rôle d'une variable dans la structure de données.

Certes, sur des données simulées, la méthode proposée permet d'atteindre les résultats attendus même en présence d'une mauvaise spécification du modèle de réponse. Cependant, il serait intéressant d'étudier plus en profondeur le comportement de cette méthode en fonction de différentes situations. En effet, le BMA DAG peut ne pas réussir à faire la distinction entre un confondant et une variable intermédiaire dans certains cas. Par exemple, considérons un DAG simple à trois variables : réponse, exposition et une troisième variable qui peut être un confondant ou une variable intermédiaire. Les deux DAG vont impliquer la même probabilité a posteriori et la méthode ne va donc pas réussir à faire la distinction entre eux.

L'approche proposée dans ce projet présuppose que le DAG de base est très bien connu et que les autres DAG considérés sont similaires. En effet, seulement une variation dans les flèches qui fait la différence entre eux. Les DAG qui sont très différents, par exemple avec des variables complètement déconnectées, n'ont pas été traités. Dans les travaux de recherche futurs, il serait utile de tester la méthode sur des structures fondamentalement différentes.

Bibliographie

Paul D. Allison. Multiple imputation for missing data : A cautionary tale. *Sociological Methods & Research*, 28(3) :301–309, 2000.

Vic Barnett and Lewis. *Outliers in statistical data*. Wiley, New York, 3rd edition, 1994.

Todd E. Bodner. What improves with increased missing data imputations ? *Structural Equation Modeling : A Multidisciplinary Journal*, 15(4) :651–675, 2008.

Bradley P. Carlin and Thomas A. Louis. *Bayesian methods for data analysis*. CRC Press, Florida, 3rd edition, 2008.

M. Clyde. *Subjective and objective Bayesian statistics*. James Press : Wiley-Interscience, New Jersey, 2nd edition, 2003.

Rhian Daniel, Jingjing Zhang, and Daniel Farewell. Making apples from oranges : Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3) :528–557, 2021.

Thierry Duchesne. STT 7140 Statistique bayésienne. *Recueil inédit, Université Laval*, 2012.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis.* Chapman & Hall/CRC, Florida, 2nd edition, 2004.

M. Maria Glymour and Sander Greenland. Causal diagrams. *Modern epidemiology*, 3 : 183–209, 2008.

Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1) :37–48, 1999a.

Sander Greenland, Judea Pearl, and James M. Robins. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1) :29–46, 1999b.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv :1907.02435*, 2019.

Miguel A. Hernán and James M. Robins. *Causal inference : What if.* Chapman and Hall/CRC, Florida, 2020.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging : a tutorial. *Statistical Science*, 14(4) :382–401, 1999.

Brandon Koch, David M. Vock, and Julian Wolfson. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74(1) :8–17, 2018.

Sang Kyu Kwak and Jong Hae Kim. Statistical data preparation : Mnagement of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4) :407, 2017.

Hongjing Liao, Yanju Li, and Gordon Brooks. Outlier impact and accommodation methods : Multiple comparisons of type I error rates. *Journal of Modern Applied Statistical Methods*, 15(1) :23, 2016.

Roderick JA Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3) :287–296, 1988.

Judea Pearl. *Causality : Models, reasoning, and inference*. Cambridge University Press, New York, 2nd edition, 2009.

Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov

equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18 (1) :8132–8193, 2017.

Adrian E. Raftery, David Madigan, and Jennifer A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92 (437) :179–191, 1997.

Craig A. Rolling and Yuhong Yang. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(4) :749–769, 2014.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1) :41–55, 1983.

Bernard Rosner. On the detection of many outliers. *Technometrics*, 17(2) :221–227, 1975.

Bernard Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2) :165–172, 1983.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv :1912.00306*, 2019.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5) :688–701, 1974.

Donald B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1) :87–94, 1986.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434) :473–489, 1996.

Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York, 2004.

Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4) :512–522, 2009.

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

Susan M. Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso : Variable selection for causal inference. *Biometrics*, 73(4) :1111–1122, 2017.

Denis Talbot, Geneviève Lefebvre, and Juli Atherton. The Bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2) :207–236, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1) :267–288, 1996.

Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3) :661–671, 2012.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations : Issues and guidance for practice. *Statistics in Medicine*, 30(4) : 377–399, 2011.

Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246) : 1–45, 2020.