UNIVERSITÉ LAVAL

Estimation bayésienne du lasso adaptatif pour l'issue

Mémoire

Serigne Abib Gaye

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

Québec, Canada

Estimation bayésienne du lasso adaptatif pour l'issue

Mémoire

Serigne Abib GAYE

Sous la direction de:

Denis Talbot, directeur de recherche

Résumé

Dans ce mémoire, on cherche à développer une nouvelle méthode d'estimation pour le lasso adaptatif pour l'issue en utilisant la machinerie bayésienne. L'hypothèse de recherche est que notre nouvelle méthode va beaucoup réduire la lourdeur computationnelle du lasso adaptatif pour l'issue.

Notre méthode utilise les mêmes fondements théoriques que le lasso adaptatif pour l'issue. Elle remplit donc les conditions de la propriété d'oracle. Pour sa mise en œuvre, on ajuste d'abord un modèle du score de propension bayésien. Ensuite, on estime l'effet du traitement moyen par la pondération par l'inverse de la probabilité de traitement. Par ailleurs, nous considérons une distribution gamma pour le paramètre de régularisation λ qui nous permet de choisir ce paramètre à partir d'un ensemble continu, alors que le lasso adaptatif pour l'issue fréquentiste utilise une approche de validation croisée qui doit faire un choix parmi un ensemble discret de valeurs pré spécifiées.

In fine, la méthode que nous avons développée répond bien à nos attentes, et permet donc de produire les inférences de façon beaucoup plus rapide. En effet, il a fallu seulement 41.298 secondes pour que cette méthode effectue les inférences, alors que 44.105 minutes ont été nécessaires au lasso adaptatif pour l'issue.

On espère que les idées développées dans ce mémoire vont contribuer significativement à améliorer les méthodes de sélection de variables en inférence causale avec l'appui des techniques bayésiennes.

Abstract

In this paper, we aim to develop a new estimation method for the outcome adaptive lasso using Bayesian machinery. The research hypothesis is that our new method will significantly reduce the computational burden of the outcome adaptive lasso.

Our method uses the same theoretical foundation as the outcome adaptive lasso. It therefore meets the oracle properties. For its implementation, Bayesian propensity score model is first fitted. Next, the average treatment effect is estimated using inverse probability of treatment weights. In addition, we consider a gamma distribution for the regularisation parameter λ in order to choose this parameter over a continuous set, whereas the frequentist outcome adaptive lasso uses a cross-validation procedure that selects λ among a prespecified discrete set.

In fine, the method we have developed meets our expectations, and therefore makes it possible to produce inferences much faster. Indeed, it took only 41.298 seconds for this method to yield inferences, while 44.105 minutes were required for the outcome adaptive lasso.

We hope that the ideas developed in this paper will significantly contribute to improve methods for selecting variables in causal inference with the support of Bayesian techniques.

Table des matières

Liste des tableaux

Remerciements

C'est avec un immense plaisir que j'adresse mes remerciements les plus vifs et les plus sincères à la personne qui a accepté de m'encadrer pour réaliser ce projet de mémoire. Cette personne a plus que répondu à mes attentes grâce à sa pédagogie, sa disponibilité mais aussi ses qualités humaines qui sont appréciables. Je veux nommer mon directeur de recherche, Dr. Denis Talbot. Je remercie aussi Monsieur Didier Brassard pour sa collaboration avec les données réelles qu'il m'a fournies. Je remercie également toutes les personnes qui de près ou de loin ont contribué à la réalisation de ce travail.

Introduction

Un problème récurrent en inférence causale est la mesure de l'effet d'une exposition ou d'un traitement dans une population. Souvent, nous avons deux groupes de traitement : un groupe traité et un groupe non traité. L'assignation du traitement dans les groupes se fait en général par la randomisation ou la non-randomisation. Les études avec assignation aléatoire sont l'idéal car elles permettent une meilleure comparabilité entre les groupes de traitement, et une meilleure évaluation de l'efficacité d'un traitement. En effet, comme les caractéristiques initiales de la population sont en moyenne identiques entre les groupes dans ces études, la seule différence observée sera attribuée à l'effet du traitement. Toutefois, les études avec assignation aléatoire sont souvent impossibles pour des raisons éthiques ou de faisabilité. Par exemple, lorsqu'on s'intéresse à l'effet d'une bonne alimentation sur la tension artérielle, il serait problématique d'exposer à un groupe une mauvaise alimentation et l'autre à une bonne alimentation.

En l'absence d'études avec assignation aléatoire, on peut avoir recours aux études observationnelles où l'assignation du traitement n'est plus sous le contrôle de l'expérimentateur. Le traitement reçu par un sujet est alors le résultat d'un processus non aléatoire et généralement inconnu ou partiellement inconnu. Dans ce contexte, il peut y avoir des déterminants communs au choix du traitement et de la variable réponse. Lorsqu'on considère toujours l'exemple de l'effet d'une bonne alimentation sur la pression artérielle, le revenu pourrait être une variable confondante, parce qu'il serait associé à une bonne alimentation et à une meilleure prise en charge sanitaire. Le score de propension (PS) qui est défini comme la probabilité pour un individu d'être exposé étant donné ses covariables, est particulièrement utile dans les études observationnelles puisqu'il permet de simuler un contexte d'étude randomisé.

Une hypothèse importante dans les modèles PS, est que tous les facteurs confondants de la relation entre le traitement et l'issue sont mesurés et inclus dans le modèle. Ajuster pour tous les facteurs potentiellement confondants est une stratégie commune, mais peut produire des estimateurs dont la variance est élevée à cause des problèmes de multicollinéarité. L'identification de facteurs confondants doit se faire à l'aide des connaissances du domaine d'application. Les connaissances sont souvent insuffisantes pour faire ce choix parfaitement. Plusieurs au-

teurs ont discuté de cette problématique d'identification de facteurs confondants, et suggéré de supplémenter les connaissances du domaine d'application par des informations provenant des données récoltées, Shortreed & Ertefaie (2017)[40], Robins & Greenland (1986)[33], Judkins et al. (2007)[21], Schneeweiss et al. (2009)[38], Vansteelandt et al. (2012)[45], Van der Laan & Gruber (2010)[44], Rolling & Yang (2013)[34], et Talbot et al. (2015a)[41]. Une méthode efficiente d'identification de facteurs confondants doit prendre en compte à la fois les relations issue-covariable et traitement-covariable (Shortreed et Ertefaie, 2017)[40]. Dans ce sens, la méthode du lasso adaptatif pour l'issue a été développée pour l'identification de facteurs confondants pour l'estimation non biaisée de l'effet d'une exposition binaire sur une issue. Cette méthode du score de propension dépend d'un paramètre de régularisation λ qui est estimé à partir des données. Les inférences sont ainsi faites en supposant λ connu ou en utilisant un type de bootstrap adapté au problème d'identification de facteurs confondants. Shortreed & Ertefaie, (2017)[40] ont utilisé 10 000 réplications, ce qui alourdit la procédure.

On retrouve dans la littérature des modèles PS, des études sur l'estimation de la variance de l'effet du traitement (Kaplan et Chen, 2012)[22]. Hirano, Imbens, & Ridder (2003)[16] ; Lunceford & Davidian (2004)[26] ont par exemple développé des estimateurs de la variance de l'effet du traitement avec les méthodes de la pondération et de l'ajustement. Une approche de bootstrap est également étudiée pour estimer la variance de l'effet du traitement (Lechner, 2002[24] ; Austin & Mamdani (2006)[4]). En outre, des travaux qui adoptent des approches bayésiennes aux méthodes d'ajustement du score de propension ont été proposés (Saarela et al., 2015)[37] : la pondération inverse par probabilité de traitement (Hoshino, 2008[18] ; Kaplan et Chen, 2012[22]), l'ajustement (McCandless, Gustafson et Austin, 2009[28] ; McCandless et al, 2010[27] ; Zigler et al, 2013[50]) et l'appariement (An, 2010)[1].

Notre objectif dans ce mémoire est de développer une méthode bayésienne d'ajustement du lasso adaptatif pour l'issue. En effet, les approches bayésiennes utilisent une spécification complète d'un modèle de probabilité et permettent de produire des inférences en tenant compte de l'ensemble du processus. On pourra donc tenir compte du choix de λ, du choix des variables et de l'estimation de l'effet causal dans une même procédure. L'hypothèse est que cette procédure sera moins exigeante d'un point de vue computationnel que le recours au bootstrap lissé afin de produire des inférences. Un avantage supplémentaire sera de pouvoir effectuer un choix de λ sur un ensemble continu, alors que le lasso adaptatif pour l'issue fréquentiste effectue un choix par validation croisée parmi un ensemble restreint proposé par l'utilisateur.

Dans ce qui suit, nous allons d'abord présenter les données sur lesquelles la méthode à développer est appliquée. Plus spécifiquement, nous cherchons à estimer l'effet de la saine alimentation sur la pression sanguine systolique. Ensuite un rappel sera fait sur l'inférence causale et la statistique bayésienne, avant de décrire les différentes méthodes du lasso. Nous

terminerons par la présentation de la méthode bayésienne d'estimation du lasso adaptatif pour l'issue.

Chapitre 1

Cadre de l'étude PREDISE

1.1 Généralités

Il existe aujourd'hui un large consensus sur le rôle important de la saine alimentation pour la santé et pour la prévention de nombreuses maladies chroniques telles que le diabète, l'obésité, les maladies cardiovasculaires (MCV), l'ostéoporose et certains types de cancers (Blanchet, Plante et Rochette, 2009)[5].

« La surveillance régulière des apports alimentaires et nutritionnels au sein de la population est donc indispensable pour des politiques de santé publiques efficaces en matière de nutrition »(Brassard et al., 2018)[6].

« L'étude PREDISE (Prédicteurs Individuels, Sociaux et Environnementaux) est une étude transversale multicentrique qui visait à recueillir au moyen d'une plateforme Web des données de nature individuelle, sociale et environnementale sur des facteurs associés à l'adhésion aux recommandations sur la saine alimentation »(Brassard et al., 2018)[6]. Cette étude a concerné la population francophone adulte dans cinq régions administratives différentes de la province de Québec : Estrie, Saguenay-Lac-Saint-Jean, Capitale-Nationale/Chaudière-Appalaches, Montréal et Mauricie, et s'est déroulée entre 2015 et 2017. L'étude a utilisé la méthode du rappel de 24 heures (R24W) dans la collecte de l'information et porte sur un échantillon de 1147 individus[6].

Dans cette étude, la qualité de l'alimentation est mesurée à travers la version canadienne du Healthy Eating Index (HEI). Le HEI est un indice composite, développé par le département de l'agriculture des États-Unis pour mesurer la qualité de l'alimentation des américains. La version originale de 1995 était basée sur les recommandations diététiques pour les Américains et la pyramide alimentaire. Il comprend deux principales composantes : « Adequacy compo-

nents » et « Moderation components », et a une valeur maximale de 100 points. La première composante regroupe les aliments plus nutritifs et ceux dont la consommation est encouragée, soit en plus grande quantité et plus fréquemment. La seconde contient les aliments les moins nutritifs et ceux dont la consommation est découragée, soit en moins grande quantité et moins fréquemment.

Suivant le niveau du score, le département américain de l'agriculture définit trois types de régime alimentaire possible : un régime alimentaire de bonne qualité pour un score supérieur à 80 points, un régime alimentaire à améliorer pour un score qui se situe entre 50 et 80 points et un mauvais régime alimentaire pour un score inférieur à 50 points (Garriguet, 2009)[14].

Ce score a connu une révision en 2005 relativement facile à adopter pour le Canada ; il s'agit du C-HEI[A]. Le C-HEI traduit les recommandations du régime alimentaire pour le Canada en 2007. L'adaptation canadienne comprend huit composantes classées dans « Adequacy components » (légumes et fruits, fruits entiers, légumes vert foncé et orangés, produits céréaliers, grains entiers, lait et substituts, viande et substituts et gras insaturés), et trois composantes dans « Moderation components » (gras saturés, sodium et autre aliment). Ces deux principales composantes du score C-HEI totalisent un maximum de 60 points et un maximum de 40 points respectivement (Garriguet, 2009)[14]. Les recommandations du guide alimentaire de 2007 du Canada sont établies dans le tableau ci-dessous.

Tableau 1.1 – Nombre de portions du Guide alimentaire recommandé chaque jour

	Adolescents		Adultes			
Âge (ans)	14-18		19-50		51+	
Sexe	Filles	garçons	Femmes	Hommes	Femmes	Hommes
Légumes et fruits	7	8	7-8	8-10	7	7
Produits céréaliers	6	7	6-7	8	6	7
Lait et substituts	3-4	3-4	2	2	3	3
Viandes et substituts	2	3	2	3	2	3

Source : www.santecanada.gc.ca/guidealimentaire

Le guide alimentaire est incontournable dans la lutte contre les maladies chroniques causées par l'alimentation telle que l'hypertension artérielle. On parle d'hypertension lorsque la pression systolique est supérieure à 140 millimètres de mercure ou lorsque la pression

A. Canadian Healthy Eating Index

diastolique est supérieure à 90 millimètres de mercure.

Schwingshackl et al. (2017)[39] ont montré dans une méta-analyse, une association inverse au risque d'hypertension avec la consommation journalière de 30g de grains entiers (RR[B] : 0.92 ; 95% CI : 0.87, 0.98), 100g de fruits (RR : 0.97 ; 95% CI : 0.96, 0.99), 28g de noix (RR : 0.70 ; 95%CI : 0.45, 1.08), et 200g de produits laitiers (RR : 0.95 ; 95% CI : 0.94, 0.97). Par contre la consommation journalière de 100g de viande rouge (RR : 1.14 ; 95% CI : 1.02, 1.28), 50g de viande transformée (RR : 1.12 ; 95% CI : 1.00, 1.26), et 250mL de boisson sucrée (RR : 1.07 ; 95% CI : 1.04, 1.10) a une association positive avec l'hypertension.

1.2 Caractéristiques de la population de l'étude

Dans cette section, nous allons analyser la répartition du score C-HEI suivant certaines caractéristiques de la population. Ce score est dichotomisé pour avoir deux types de régime alimentaires ; un mauvais régime pour un score C-HEI inférieur à 50, et un bon régime alimentaire pour un score supérieur ou égal à 50.

La différence de moyenne standardisée (SMD) est ici utilisée pour comparer la distribution des covariables entre les groupes. Austin (2009)[2] indique qu'une SMD supérieure à 0.1 reflète une différence importante entre les groupes de traitement de la distribution des covariables.

Le score moyen global qui est de l'ordre de 54.5, dénote un faible niveau d'adhésion aux recommandations du guide alimentaire de la population adulte francophone dans les cinq régions administratives de la province de Québec (Brassard et al., 2018)[6]. Par ailleurs, on constate que 36.2% de la population ayant une saine alimentation est constituée de personnes âgées entre 50 et 65 ans. Ce groupe de personne a un effectif absolu de 264.7. La saine alimentation est aussi plus présente chez les femmes (58.3%), chez les personnes ayant un niveau d'étude universitaire (49.9%), chez la population la plus riche (35.7%), chez les non fumeurs (56.6%), dans les centres INAF (38.0%) et IRCM (36.1%), dans la population avec une masse corporelle normale (43.1%), et dans la population ayant une forte connaissance informatique (54.5%).

B. Risque relatif

Tableau 1.2 – Répartition du score C-HEI

Caractéristiques [C]	<50	≥50	SMD
n	415.9	731.1	
Groupe d'âge (%)			0.131
18-34	164.7 (39.6)	243.6 (33.3)	
35-49	115.6 (27.8)	222.8 (30.5)	
50-65	135.6 (32.6)	264.7 (36.2)	
Sexe = femme (%)	149.5 (35.9)	426.3 (58.3)	0.460
Education (%)			0.366
Secondaire ou moins	129.6 (33.9)	139.5 (19.7)	
CEGEP	119.8 (31.3)	215.4 (30.4)	
Université	133.0 (34.8)	353.4 (49.9)	
Revenu,CAD (%)			0.109
<30000	61.4 (18.0)	100.1 (15.3)	
30000 à < 60000	101.1 (29.7)	182.9 (28.0)	
60000 à < 90000	59.4 (17.4)	136.7 (20.9)	
≥ 90000	118.9 (34.9)	233.2 (35.7)	
Ethnicité (%)			0.075
Caucasien	360.4 (95.2)	642.5 (93.8)	
Africain-américain	8.4 (2.2)	16.8 (2.5)	
Hispanique	5.8 (1.5)	12.3 (1.8)	
Autre	4.0 (1.1)	13.0 (1.9)	
Centre (%)			0.117
CHUS	42.3 (10.2)	67.8 (9.3)	
ECOGENE-Ch	41.8 (10.0)	64.9 (8.9)	
INAF	156.7 (37.7)	278.0 (38.0)	
IRCM	133.0 (32.0)	263.9 (36.1)	
UQTR	42.1 (10.1)	56.6 (7.7)	
Indice de masse corporelle (%)			0.249
Normale(<25)	114.1 (32.0)	287.3 (43.1)	
Surpoids(25-30)	124.2 (34.9)	217.7 (32.6)	
Obèse(>30)	118.0 (33.1)	162.3 (24.3)	
Statut fumeur (%)			0.445
Oui	101.6 (24.4)	61.1 (8.4)	
Autrefois	122.8 (29.5)	256.0 (35.0)	
Jamais	191.4 (46.0)	414.0 (56.6)	
Régime = régime (%)	27.1 (6.5)	94.4 (12.9)	0.217
Complément alimentaire = Oui (%)	77.9 (18.7)	225.0 (30.8)	0.282
Connaissances informatiques (%)			0.220
Faible	40.4 (10.6)	38.6 (5.5)	
Moyen	124.9 (32.7)	280.5 (40.0)	
Fort	217.1 (56.8)	382.0 (54.5)	
médication = non (%)	161.6 (42.2)	330.7 (47.1)	0.098
Activité physique par semaine (mean (SD))	2.70 (2.75)	3.51 (2.86)	0.287
Biais de désirabilité (mean (SD))	11.70 (4.52)	11.82 (4.48)	0.025

C. Le nombre et le pourcentage (entre parenthèses) sont rapportés pour les variables catégorielles.
Nous avons également omis les données manquantes dans notre analyse.
Les nombres ne sont pas entiers en raison de la pondération appliquée aux données, pour assurer la représentativité selon l'âge et le sexe dans chaque région (voir Brassard, 2018 [6]).

Chapitre 2

RAPPELS

2.1 Rappel sur l'inférence causale

Le but premier de l'inférence causale est d'établir une relation de cause à effet non biaisée. Autrement dit, il s'agit de démontrer que certains éléments sont les causes des effets que nous observons.

Soit A une variable d'exposition ou de traitement dichotomique et Y, une variable de réponse ou d'issue. Le traitement A a un effet causal sur l'issue Y pour un individu i si $Y_i^{a=1} \neq Y_i^{a=0}$ (Hernán & Robins, 2020)[15]. Les variables $Y^{a=1}$ et $Y^{a=0}$ sont appelées des issues potentielles ou contrefactuelles et $Y_i^{a=1}$ et $Y_i^{a=0}$ désignent respectivement la réponse sous le traitement et la réponse sous le contrôle pour l'individu i. Par ailleurs, pour chaque individu, c'est uniquement la réponse sous le traitement que l'individu a réellement reçu qui est observée (Hernán & Robins, 2020)[15]. Il n'est donc généralement pas possible d'identifier pour tous les individus l'effet causal. En contrepartie, on peut définir l'effet du traitement moyen (ATE) dans la population des individus : $E[Y^{a=1}]$ - $E[Y^{a=0}]$. E(.) représente la moyenne dans la population.

Hernán et Robins établissent également trois mesures de l'effet causal à savoir la différence de risque contrefactuelle (i), le risque ratio (ii) et le rapport de cotes (iii).

(i) $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$

(ii) $\dfrac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]}$

(iii) $\dfrac{\Pr[Y^{a=1} = 1]/\Pr[Y^{a=1} = 0]}{\Pr[Y^{a=0} = 1]/\Pr[Y^{a=0} = 0]}$

2.1.1 Score de propension

Le score de propension correspond à la probabilité pour un individu d'être exposé à un traitement étant donné ses covariables (Rosenbaum et Rubin, 1983)[36] e(X)=Pr(A=1| X)=E(A|X), où X désigne un ensemble de covariables préexposition potentiellement confondantes. Ces variables sont conceptualisées comme déterminantes à la fois de l'exposition A et de la réponse Y.

La validité des modèles PS pour produire une estimation non biaisée de l'effet du traitement repose sur trois conditions. La condition de positivité assure que la probabilité de recevoir chaque valeur du traitement est strictement supérieure à zéro, $0 < \Pr(A=1|X) < 1$. La condition d'échangeabilité s'exprime mathématiquement par : $Y^a \perp\!\!\!\perp A|X$, et traduit que la probabilité conditionnelle de recevoir chaque valeur de traitement dépend seulement des covariables mesurées. La condition de cohérence indique que pour un traitement a, assigné au sujet i, l'issue observée pour ce sujet correspond à son issue contrefactuelle associée au traitement a : $A_i = a \Rightarrow Y_i = Y_i^a$ (Hernán & Robins, 2020)[15].

L'idée du score de propension est d'assurer un équilibre dans la distribution des covariables préexposition entre les sujets exposés et non exposés. Ainsi, pour un score de propension donné e(x), il existe une indépendance statistique entre l'exposition et les covariables X; $A \perp\!\!\!\perp X|e(x)$.

Démonstration
La démonstration de l'indépendance statistique soulignée tantôt est inspirée de Diop, 2019[8]. Elle revient à prouver l'égalité suivante :

$$E(A|X, e(X)) = E(A|e(X)). \tag{2.1}$$

Nous allons démontrer que le terme de gauche et celui de droite de l'équation 2.1 sont égaux au score de propension.

e(X) est une fonction surjective de X. Autrement dit, à chaque élément de X, on associe une et une seule valeur e(x), mais la réciproque est fausse : une même valeur e(x) peut correspondre à différentes valeurs des covariables X. Ainsi, e(X) ne contient aucune information qui ne soit pas déjà contenue dans X. Le conditionnement par rapport à X et e(X) peut donc se simplifier au conditionnement par rapport à X seulement. Alors, on a que E(A|X, e(X))=E(A|X) qui est égale au score de propension par définition.

Le calcul du membre de droite de l'équation 2.1 se base sur le théorème de l'espérance totale.

$$
\begin{aligned}
E(A|e(X)) &= E\{E(A|X, e(X))|e(X)\} \\
&= E(E(A|X)|e(X)) \\
&= E(e(X)|e(X)) \\
&= e(X) \blacksquare
\end{aligned}
$$

La propriété $A \perp\!\!\!\perp X|e(x)$ indique que le score de propension permet de simuler un contexte d'étude randomisée conditionnelle puisque dans les expériences randomisées, les caractéristiques initiales des groupes exposés et non exposés sont en moyenne identiques.

Des méthodes d'ajustement basées sur le score de propension sont proposées dans la littérature notamment, l'appariement, la pondération et l'ajustement à l'aide de données observationnelles.

Pondération :
Les méthodes de pondération du score de propension trouvent leur fondement dans l'idée d'échantillonnage de Horvitz-Thompson (Horvitz et Thompson, 1952)[17], et visent à pondérer les sujets exposés et les sujets non exposés en fonction de leur score de propension (Kaplan et Chen, 2012)[22].

L'inverse de la probabilité de traitement est utilisée pour estimer l'effet du traitement moyen. Le poids est défini par : $\frac{A}{e(x)} + \frac{1-A}{1-e(x)}$, et correspond pour chaque sujet à l'inverse de la probabilité de recevoir le traitement que le sujet a vraiment reçu. A=1 pour un sujet exposé et 0 sinon.

2.1.2 IPTW

La pondération par la probabilité inverse du traitement (IPTW) donne une population synthétique dans laquelle le traitement est indépendant des covariables pré-traitement (Austin, 2010)[3]. Lunceford & Davidian (2004)[26] ont décrit plusieurs estimateurs de l'effet du traitement utilisant la pondération par la probabilité inverse de traitement dont nous présentons deux estimateurs dans la suite (preuve voir Lunceford & Davidian (2004)[26]). Le premier estimateur IPTW de la différence de risque proposé par Rosenbaum (1987)[35] est :

$$\hat{\Delta}_{IPTW} = \frac{1}{N}\sum_{i=1}^{N}\frac{A_i Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-A_i)Y_i}{1-\hat{e}_i},$$

N désigne le nombre de sujets dans l'échantillon, et $\hat{e}_i$ correspond au score de propension estimé pour le sujet i.

Le second estimateur requiert de spécifier un modèle du score de propension et un modèle de régression qui lie l'issue au traitement et aux covariables pré-traitement. Soit

$m_a(X,\alpha_a)$=E(Y|A=a, X). Alors

$$\hat{\Delta}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\frac{A_iY_i - (A_i - \hat{e}_i)m_1(X_i,\hat{\alpha})}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-A_i)Y_i - (A_i - \hat{e}_i)m_0(X_i,\hat{\alpha})}{1-\hat{e}_i},$$

$\hat{\Delta}_{DR}$ posséde la propriété de "double robustesse" dans le sens que l'estimateur demeure cohérent si soit (i) le modèle du score de propension est correctement spécifié mais que les deux modèles de régression de l'issue m_0 et m_1 ne le sont pas ou (ii) si m_0 et m_1 sont correctement spécifiés mais que le modèle du score de propension ne l'est pas (Lunceford et Davidian, 2004)[26].

2.2 Rappel sur la statistique bayésienne

2.2.1 Principes de base de la statistique bayésienne

La statistique bayésienne, tout comme la statistique fréquentiste utilise un modèle liant les données, disons D au paramètre θ. Dans l'approche fréquentiste, l'estimation des paramètres est basée sur la fonction de vraisemblance. Des auteurs comme Robert et Casella (2010)[32] soulignent quelques limites des méthodes du maximum de vraisemblance dont la présence de plusieurs modes.

L'approche bayésienne quant à elle, résulte d'un problème d'intégration et se base sur une loi *a priori*, disons $\pi(\theta)$ qu'il faut spécifier avant de collecter les données, pour approcher les valeurs possibles de θ (Duchesne, 2012)[9]. Connaissant le modèle f(D|θ) et la loi *a priori*, la distribution *a posteriori* de θ est spécifiée suivant la formule de Bayes :

$$\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}. \tag{2.2}$$

D'autres paramètres différents de θ peuvent s'associer à la loi *a priori*. Ces paramètres qu'on peut noter par ϕ sont appelés des hyper-paramètres dont les valeurs peuvent être connues ou pas. Si la valeur de ces hyper-paramètres n'est pas connue, alors on ajoute un niveau à la hiérarchie : D|θ ~ L, $\theta|\phi$ ~ L' (L et L' sont des fonctions de vraisemblance des modèles f(D|θ) et f($\theta|\phi$)), et on suppose une loi pour ϕ, disons $\pi(\phi)$. Nous avons dans ce cas une loi *a priori* $\pi(\theta|\phi)$, et la loi *a posteriori* de θ se calcule en appliquant la règle des probabilités totales au numérateur et au dénominateur de (2.2) :

$$\pi(\theta|D) = \frac{\int f(D|\theta)\pi(\theta|\phi)\pi(\phi)d\phi}{\iint f(D|\theta)\pi(\theta|\phi)\pi(\phi)d\phi d\theta}. \tag{2.3}$$

D'autres niveaux peuvent également être ajoutés à la hiérarchie ; lorsque par exemple la loi *a priori* $\pi(\phi|\lambda)$ dépend de paramètres λ qui suivent eux-mêmes une loi *a priori* $\pi(\lambda)$

(Duchesne, 2012)[9].

Une alternative à la modélisation hiérarchique, appelée méthode de "Bayes empirique", consiste à estimer ϕ à partir de la loi "marginale" de D sachant ϕ,

$$m(D|\phi) = \int f(D|\theta)\pi(\theta|\phi)d\theta,$$

en utilisant par exemple $\hat{\phi} = \underset{\phi}{argmax}\ m(D|\phi)$ ou une méthode des moments, et à faire ensuite des inférences en utilisant $\pi(\theta|\hat{\phi})$ comme loi *a priori* et en corrigeant les inférences pour tenir compte de la variabilité dans $\hat{\phi}$ (Duchesne, 2012)[9].

Lois a priori :
La définition d'une loi *a priori* est souvent une étape difficile pour l'analyste puisqu'elle ne dépend pas d'un algorithme précis. Il existe toutefois trois méthodes simples pour définir cette loi *a priori*.

"Élicitation" de la loi *a priori* consiste à définir des valeurs plausibles du paramètre en se basant sur l'opinion d'"experts" dans le domaine.

Une loi *a priori* est dite non informative lorsqu'elle ne permet pas de renseigner *a priori* sur les valeurs possibles du paramètre. Les lois *a priori* non informatives sont souvent des distributions impropres, c'est-à-dire qu'elles ne définissent souvent pas de vraies densités de probabilité. La loi *a priori* de Jeffreys est un exemple de loi non informative qui permet de placer une distribution vague (impropre) sur l'ensemble des paramètres conjointement (Jeffreys, 1946)[20],

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

où $I(\theta)$ est le déterminant de l'information de Fisher,

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 log(f(D|\theta))}{\partial\theta^2}\right].$$

Les lois conjuguées se démarquent de par leurs propriétés mathématiques qui rendent simples les calculs des lois *a posteriori* et prédictives. Une astuce consiste à regarder la vraisemblance f(D|θ) comme fonction de θ et de prendre une loi *a priori* "de la même forme" que la vraisemblance (Duchesne, 2012)[9].

2.2.2 Les méthodes de chaîne de Markov Monte Carlo (MCMC)

On se rappelle que toute inférence bayésienne est basée sur la loi *a posteriori*. Plusieurs options s'offrent à l'analyste pour évaluer numériquement cette loi *a posteriori*, $\pi(\theta|D)$:

- utiliser une méthode d'intégration numérique (e.g., quadrature) pour évaluer l'intégrale ;
- approximer la valeur de $\pi(\theta|D)$;
- simuler des réalisations de $\pi(\theta|D)$.

La simulation de nombres aléatoires permet souvent d'obtenir les propriétés requises des lois *a posteriori* et prédictives en inférence bayésienne, grâce notamment à la facilité et la rapidité avec laquelle les ordinateurs peuvent générer ces nombres aléatoires (Duchesne, 2012)[9].

Notre objectif principal en statistique bayésienne sera de simuler des réalisations $\theta_1, ..., \theta_N$, à partir de $\pi(\theta|D)$ afin de calculer des moments ou des quantiles de la distribution *a posteriori*. Pour ce faire, on peut avoir recours aux *méthodes non itératives* qui consistent à génèrer directement un seul échantillon θ_1,..., θ_N. Sinon, avec les méthodes de Chaîne de Markov Monte Carlo (MCMC), on peut simuler une valeur de θ à la fois, avec θ_k qui dépend des valeurs de θ_0,..., θ_{k-1} (Duchesne, 2012)[9].

Chaînes de Markov :
Une chaîne de Markov est un processus aléatoire tel que tous les vecteurs ont le même domaine (*espaces d'états*) et tel que la loi conditionnelle du vecteur aléatoire $\mathbf{x}^{(k)}$ sachant la valeur des vecteurs aléatoires $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k-1)}$ ne dépend que de la valeur de $\mathbf{x}^{(k-1)}$. Un processus aléatoire est quant à lui défini comme une suite indicée $\{\mathbf{x}^{(l)}, l = 0, 1, ..., \} = \{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...\}$ de variables/vecteurs aléatoires (Duchesne, 2012)[9]. La fonction $\phi(s,t)$ qui donne la valeur de la probabilité (espace d'états discret) ou de la densité (espace d'états continu) $p(x^{(k)} = t|x^{(k-1)} = s)$ est appelé *noyau de transition* de la chaîne.« La propriété fondamentale des chaînes de Markov est que sous certaines conditions de régularité pour l'espace d'états et le noyau de transition, on a que $\lim_{k\to\infty} p(\mathbf{x}^{(k)} = \mathbf{s}) = p^*(\mathbf{s})$ pour tout s dans l'espace d'états et que si $p(\mathbf{x}^{(k-1)} = \mathbf{s}) = p^*(\mathbf{s})$, alors $p(\mathbf{x}^{(k)} = \mathbf{t}) = p^*(\mathbf{t})$. On dit que p^* est la distribution stationnaire de la chaîne car ne dépendant pas de k » (Duchesne, 2012)[9]. Lorsqu'on simule ainsi un grand nombre de réalisations $\mathbf{x}^1, ..., \mathbf{x}^B, \mathbf{x}^{B+1}, ..., \mathbf{x}^{B+N}$, la distribution de chaque $\mathbf{x}^k$ pour $k \in \{B+1, B+2, ..., B+N\}$ sera approximativement p^*, pour B suffisamment grand (voir livre de Robert et Casella, 2010[32]).

Après avoir fixé une valeur de départ arbitraire $\theta^{(0)}$, on peut utiliser le noyau de la chaîne pour simuler des réalisations $\theta^{(1)}, ..., \theta^{(B)}, \theta^{(B+1)}, ..., \theta^{(B+N)}$. L'échantillon final désiré est obtenu avec les valeurs de $\theta^{(B+1)}, ..., \theta^{(B+N)}$ simulées, et donc en effaçant les B premières valeurs $\theta^{(1)}, ..., \theta^{(B)}$ (période de chauffe ou "burn-in"). Les N dernières valeurs conservées sont auto-corrélées, ce qui rend difficile la mesure de l'erreur d'estimation. Dans la situation où on s'intéresse à cette dernière, on ne garde alors qu'une valeur simulée sur, disons 5 ou 10 ("thinning"), par

exemple, $\theta^{(B+1)}, \theta^{(B+6)}$, ..., afin d'obtenir un échantillon final où la corrélation entre les valeurs est négligeable (Duchesne, 2012)[9].

Échantillonneur de Gibbs :
« L'échantillonneur de Gibbs est une méthode populaire pour les analyses MCMC. Il constitue une bonne manière d'échantillonner les distributions combinées de plusieurs variables, en se basant sur la notion suivante : pour échantillonner une distribution combinée, échantillonnez de manière répétitive à partir de ses conditions uni-dimensionnelles tout en sachant ce que vous avez vu jusqu'alors » [A].

Supposons $\theta=(\theta_1, ..., \theta_d)^T$. Les lois *a posteriori* conditionnelles complètes sont $\pi(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d, D)$, j=1,...,d. Soit $\theta^{(0)}=(\theta_1^{(0)}, ..., \theta_d^{(0)})^T$, une valeur initiale arbitraire. Alors *l'échantillonneur de Gibbs* est défini par l'algorithme suivant (Duchesne, 2012)[9] :

Étape 1 : Générer une réalisation $\theta_1^{(1)}$ de la loi *a posteriori* conditionnelle

$$\pi(\theta_1^{(1)}|\theta_2^{(0)}, ..., \theta_d^{(0)}, D).$$

Étape 2 : Générer une réalisation $\theta_2^{(1)}$ de la loi *a posteriori* conditionnelle

$$\pi(\theta_2^{(1)}|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_d^{(0)}, D).$$

.
.
.

Étape d : Générer une réalisation $\theta_d^{(1)}$ de la loi *a posteriori* conditionnelle

$$\pi(\theta_d^{(1)}|\theta_1^{(1)}, ..., \theta_{d-1}^{(1)}, D).$$

Étape d+1 : Générer une réalisation $\theta_1^{(2)}$ de la loi *a posteriori* conditionnelle

$$\pi(\theta_1^{(2)}|\theta_2^{(1)}, ..., \theta_d^{(1)}, D).$$

.
.
.

Étape (B+N)d : Générer une réalisation $\theta_d^{((B+N)d)}$ de la loi *a posteriori* conditionnelle

$$\pi(\theta_d^{(B+N)}|\theta_1^{(B+N-1)}, ..., \theta_{d-1}^{(B+N-1)}, D).$$

A. https://www.statsoft.fr/concepts-statistiques/glossaire/e/echantillonneur-gibbs.html

Chapitre 3

Le lasso et ses variantes

Dans ce chapitre, nous allons présenter les différentes méthodes du lasso en statistique fréquentiste et en statistique bayésienne. Nous allons également présenter des exemples de code R avec ces différentes méthodes.

3.1 Lasso ordinaire

La description de cette méthode de lasso ordinaire est basée sur l'article intitulé Regression Shrinkage and Selection via the Lasso (1996) de Robert Tibshirani. L'auteur y souligne deux raisons pour lesquelles l'estimation des coefficients de la régression par la méthode des moindres carrés ordinaires n'est pas toujours satisfaisante.

- La précision de la prédiction : les coefficients obtenus par la méthode des moindres carrés ordinaires, sont non biaisés mais ont une grande variance. La précision de la prédiction peut être améliorée en réduisant ou en mettant à zéro certains coefficients. Cette amélioration de la précision implique une réduction de la variance, au sacrifice de l'introduction d'un peu de biais.
- L'interprétation : L'interprétation d'un modèle contenant un grand nombre de prédicteurs peut être difficile. En identifiant un sous-ensemble plus petit de prédicteurs non nuls, l'interprétation du modèle s'en trouve facilitée.

L'auteur propose une nouvelle technique de sélection de variables appelée le lasso : "least absolute shrinkage and selection operator". Le lasso est une méthode de régularisation qui contracte la valeur des coefficients de telle sorte que certains prennent la valeur zéro. De façon concrète, le lasso minimise la somme des carrés résiduels sous la contrainte que la somme de la valeur absolue des coefficients de régression soit inférieure à une constante. Les méthodes de régularisation partent du principe que si des variables inutiles sont incluses dans le modèle,

on aura tendance à avoir des estimés $(\hat{\beta}_j)$ qui prennent des valeurs extrêmes et la variance des estimateurs des coefficients de régression peut augmenter, voire exploser (Duchesne, 2018)[10].

3.1.1 Estimation des paramètres

Soit $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, les estimations des coefficients du lasso sont définies par :

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{arg\,min} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j X_{ij} \right)^2 \; sous\ la\ contrainte \; \sum_j |\beta_j| \leq \lambda, \tag{3.1}$$

où λ est un paramètre de régularisation. La solution pour tout λ de α est $\hat{\alpha}$=$\bar{y}$. On admet sans perte de généralité que $\bar{y}$=0 et on omet α. On suppose aussi que les X_j sont centrées en 0.

Fixons $\lambda \geq 0$. Le problème (3.1) peut être exprimé comme un problème des moindres carrés avec 2^p contraintes d'inégalités, correspondant aux 2^p différents possibles signes de β_j. Par exemple, si $p = 2$, la contrainte $\sum_j |\beta_j| \leq \lambda$ s'exprime de façon équivalente sous la forme des quatres inégalités linéaires suivantes : $\beta_1 + \beta_2 \leq \lambda; \beta_1 - \beta_2 \leq \lambda; -\beta_1 + \beta_2 \leq \lambda; -\beta_1 - \beta_2 \leq \lambda$.
Lawson et Hansen (1974)[23] ont développé une procédure qui résout le problème des moindres carrés soumis à une contrainte d'inégalité linéaire générale $G\beta \leq h$. Ici, G est une matrice $m \times p$, correspondant à m contraintes d'inégalité linéaire. Le problème peut être résolu en introduisant les contraintes d'inégalité de manière séquentielle, et en cherchant une solution réalisable satisfaisant aux conditions dites de Kuhn-Tucker (Lawson et Hansen, 1974)[23].

Nous commençons par décrire l'algorithme de façon théorique, puis nous illustrons son fonctionnement par un exemple numérique simple. Soit g(β)=$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$, et soit δ_i, i=1,2,...,2^p, les p-tuples de la forme ($\pm$1, $\pm$1,...,$\pm$1). Alors la condition $\sum_j |\beta_j| \leq \lambda$ est équivalente à $\delta_i^T \beta \leq \lambda$ pour tout i. Considérons β, E=$\{$i : $\delta_i^T \beta = \lambda\}$ et S=$\{$i : $\delta_i^T \beta < \lambda$ $\}$. Notons par G_E la matrice pour laquelle les lignes sont δ_i pour tout i dans E. Soit $\mathbf{1}$ le vecteur constitué des 1 de longueur égale au nombre de ligne de G_E. L'algorithme suivant commence par E=$\{i_0\}$ où δ_{i0}=signe($\hat{\beta}$), $\hat{\beta}$ étant l'estimation globale des moindres carrés. Il résout le problème des moindres carrés sous contrainte que $\delta_{i0}^T \beta \leq \lambda$ puis vérifie si $\sum_j |\beta_j| \leq \lambda$. Si oui, le calcul est complet. Sinon, la contrainte violée est ajoutée à E et la procédure continue jusqu'à ce que $\sum_j |\beta_j| \leq \lambda$. On décline ici l'algorithme.

1. Commencer par E=$\{i_0\}$ où δ_{i0}=signe($\hat{\beta}$), $\hat{\beta}$ étant l'estimation globale des moindres carrés.
2. Trouver $\hat{\beta}$ qui minimise g(β) sous contrainte $G_E\beta \leq \lambda\mathbf{1}$.
3. Lorsque $\sum_j |\hat{\beta}_j| > \lambda$.
4. On ajoute i à l'ensemble E où δ_i=signe($\hat{\beta}$). Trouver $\hat{\beta}$ qui minimise g(β) sous contrainte $G_E\beta \leq \lambda\mathbf{1}$.

3.1.2 Exemple 1 (Lasso)

Dans cet exemple, nous allons implanter l'algorithme défini précédemment.

```
Posons lambda=0.5
On simule 100 observations suivant une loi normale pour chacune des variables
X1simul, X2simul et Ysimul.
set.seed(232474)
X1simul=rnorm(100)
X2simul=rnorm(100)
Ysimul=X1simul + rnorm(100)
***Standardisons les variables
stand.X1simul=scale(X1simul)
stand.X2simul=scale(X2simul)
stand.Ysimul=scale(Ysimul, scale = FALSE)
```

D'abord, nous allons ajuster un modèle de régression linéaire pour trouver notre δ_{i0}.

```
modeleEx=lm(stand.Ysimul ~ -1 + stand.X1simul + stand.X2simul)
On recupère le signe des coefficients du modèle ajusté.
Deltai0=as.numeric(sign(coef(modeleEx)))
Deltai0
# 1 1
Ensuite, on passe à l'étape 2 pour définir la fonction à optimiser sous
contrainte.
fr.optim1 <- function(betA){
  beta1<-betA[1]
  beta2<-betA[2]
  sum((stand.Ysimul-beta1*stand.X1simul-beta2*stand.X2simul)^2)
}
GE=matrix(Deltai0,1,2)
# GE = (1 1)
initial<-c(0,0)
lasso.Ex<-constrOptim(initial, fr.optim1, NULL, -GE, -lambda)
Les coefficients estimés du lasso sont
lasso.Ex$par
# 0.741799 -0.241799
```

A la troisième étape, on vérifie que $\sum|\beta|$=0.9835981 est supérieure à λ. On devrait donc ajouter le i correspondant aux signes des nouveaux β, c'est à dire qu'on ajoute la ligne (1 -1)

à G_E. Donc G_E est maintenant $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Puis, on reprend l'algorithme.

Toutefois, la formulation la plus usuelle de l'estimation du lasso est la suivante :

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda' \sum_{j=1}^{p} |\beta_j|. \tag{3.2}$$

On peut résoudre cette dernière de la manière dont le fait la fonction glmnet du package glmnet dans le logiciel R. Voir exemple ci-dessous.

3.1.3 Exemple 2 (Lasso)

```
fr.optim2<-function(betA) {
beta1<-betA[1]
beta2<-betA[2]
sum(0.5*(stand.Ysimul-beta1*stand.X1simul-beta2*stand.X2simul)^2)/
  length(stand.Ysimul) + lambda*(abs(beta1)+abs(beta2))
}
initial<-c(0,0)
lasso.Ex<-optim(initial, fr.optim2)
lasso.Ex$par
```

On rapelle que $\lambda' \neq \lambda$.

3.1.4 Erreur standard

« Comme l'estimation du lasso est une fonction non linéaire et non différentiable des valeurs de réponse, même pour une valeur fixe de λ, il est difficile d'obtenir une estimation précise de son erreur type. Une approche est via le bootstrap : soit λ peut être corrigé, soit optimisé pour chaque échantillon bootstrap. Corriger λ revient à sélectionner un meilleur sous-ensemble, puis à utiliser l'erreur standard des moindres carrés pour ce sous-ensemble. Une estimation approximative peut être obtenue en écrivant la pénalité $\sum_j |\beta_j|$ comme suit $\sum_j \beta_j^2/|\beta_j|$. Ainsi, pour l'estimation du lasso $\hat{\beta}$, on peut trouver une solution approximative par la régression ridge de la forme $\beta^* = (X^T X + \lambda_0 W^{-1})^{-1} X^T y$, où W est une matrice diagonale dont les éléments sont $|\hat{\beta}_j|$, W^- désigne une matrice inverse généralisée et λ_0 est choisie de telle sorte qu'on ait $\sum|\beta_j|^*=\lambda$. La matrice de covariance peut être approximée par $(X^T X + \lambda_0 W^-)^{-1} X^T X (X^T X + \lambda_0 W^-)^{-1} \hat{\sigma}^2$ [A], où $\hat{\sigma}^2$ est l'estimation de la variance de l'erreur. Cette approximation suggère un algorithme itératif de la régression ridge pour l'estimation

A. La matrice de covariance est incorrectement indiquée comme étant égale à $(X^T X + \lambda_0 W^-)^{-1} X^T X (X^T X + \lambda_0 X^-)^{-1} \hat{\sigma}^2$ dans Tibshirani (1996).

du lasso, mais cela s'avère assez inefficace. Cependant, elle est utile pour la sélection du paramètre λ »$\left[traduction\ \ libre^{[43]}\right]$.

Bien que le lasso soit un bon outil de sélection dans un problème de régression multiple, certaines limites de cette méthode sont à relever.

- Les fortes corrélations : si des variables sont fortement corrélées entre elles et qu'elles sont importantes pour la prédiction, le lasso en privilégiera une au détriment des autres. Dans ce cas, la convergence de la sélection du lasso n'est plus assurée (Zhao et Yu, 2006)[49].
- La très grande dimension : lorsque notamment la dimension est trop élevée (p très grand par rapport à N), le lasso ne sélectionne que n variables prédictives au maximum (Wainwright et Yu, 2009)[46].
- Le lasso produit des estimations biaisées pour les grands coefficients, ce qui le rend non optimal pour l'estimation (Fan et Li, 2001)[12].
- La valeur optimale de λ pour la prédiction ne permet pas d'avoir une sélection convergente de variables (Meinshausen et Bühlmann, 2004)[29]. C'est à dire que la probabilité que le modèle identifie correctement les β_j qui font partie des coefficicents nuls et ceux qui font partie des coefficients non nuls, ne tend pas vers 1 quand N tend vers l'infini.

3.2 Lasso adaptatif

Le lasso adaptatif est une extension du lasso ordinaire (Tibshirani, 1996)[43] qui utilise des coefficients spécifiques de poids (Zou, 2006)[51]. Cette méthode est définie en deux étapes. Dans un premier temps, on calcule un estimateur préliminaire $\hat{\beta} \in \mathbb{R}^N$, pouvant être l'estimateur des moindres carrés ordinaires, du lasso ou tout autre estimateur. Dans un second temps, cet estimateur préliminaire est utilisé pour ajuster la pénalité imposée sur chacun des coefficients du paramètre de régression du lasso adaptatif. Soit $\gamma>0$, et $\hat{w} = 1/|\hat{\beta}|^{\gamma}$, le vecteur poids. L'estimateur du lasso adaptatif $\hat{\beta}^{*(n)}$ est donné par (Zou, 2006)[51] :

$$\hat{\beta}^{*(n)} = \underset{\beta}{arg\,min} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|$$

3.2.1 Estimation

L'estimation des paramètres du lasso adaptatif peut être faite à l'aide de l'algorithme LARS, défini par Efron et al. 2004 comme suit (Zou, 2006)[51] :

1. définir $x_j^{**} = x_j / \hat{w}_j$, j=1,2,...,p,

2. résoudre le problème du lasso pour tout λ_n,

$$\hat{\beta}^{**} = arg\min_{\beta} \|y - \sum_{j=1}^{p} x_j^{**}\beta_j\|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

3. calculer $\hat{\beta}_j^{*(n)} = \hat{\beta}_j^{**}/\hat{w}_j$, j=1,2,...,p.

Etant donné un estimateur $\hat{\beta}$ utilisé pour définir le vecteur poids, on cherche un couple optimal (γ,λ_n) avec la méthode de la validation croisée à deux dimensions. Cependant, si l'on connait la valeur de γ, on cherche alors uniquement la valeur de λ_n optimale par validation croisée.

3.2.2 Exemple 3 (Lasso adaptatif)

Posons $gamma = 2$. On définit d'abord x_j^{**}.

```
wj = as.numeric(1/abs(coef(modeleEx))^gamma)
X.sstarj<-cbind(stand.X1simul/wj[1],stand.X2simul/wj[2])
Ensuite, on résout le problème du lasso.
fr.optim3<-function(betA){
  beta1<-betA[1]
  beta2<-betA[2]
  sum(0.5*(stand.Ysimul-beta1*X.sstarj[,1]-beta2*X.sstarj[,2])^2)/
    length(stand.Ysimul) + lambda*(abs(beta1)+abs(beta2))
}
initial<-c(0,0)
lasso.Ex.ada<-optim(initial, fr.optim3)
```

Enfin, on calcule $\hat{\beta}_j^{*(n)}$.

```
beta.starn=lasso.Ex.ada$par/wj
# 5.719200e-01 -6.280199e-11
```

3.2.3 Erreur standard

Fan et Li (2001)[12] proposent une approximation quadratique locale (LQA) débouchant sur une formule sandwich pour calculer la covariance des estimations pénalisées des composantes non nulles. Ainsi, pour des coefficients β_j non nuls, les pénalités du lasso adaptatif vérifient :

$$|\beta_j|\hat{w}_j \approx |\beta_{j0}|\hat{w}_j + \frac{1}{2}\frac{\hat{w}_j}{|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2).$$

Supposons sans perte de généralité que les d premières composantes de β sont non nulles et posons $\sum(\beta) = diag(\frac{\hat{w}_1}{\beta_1}, ..., \frac{\hat{w}_d}{\beta_d})$. Soit $\mathbf{X}_d$ la matrice comportant les d premières colonnes de $\mathbf{X}$. Le lasso adaptatif peut être estimé à travers un calcul d'itération de la régression ridge (Fan et Li, 2001)[12],

$$(\beta_1, ..., \beta_d)^T = \left(\mathbf{X}_d^T \mathbf{X}_d + \lambda_n \sum(\beta_0)\right)^{-1} \mathbf{X}_d^T \mathbf{y},$$

lequel fournit l'estimation de la matrice de covariances des composantes non nulles du lasso adaptatif $\cap_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0\}$,

$$c\hat{o}v(\hat{\beta}_{\cap_n^*}^{*(n)}) = \sigma^2 \left(X_{\cap_n^*}^T X_{\cap_n^*} + \lambda_n \sum(\hat{\beta}_{\cap_n^*}^{*(n)})\right)^{-1} \times X_{\cap_n^*}^T X_{\cap_n^*} \times \left(X_{\cap_n^*}^T X_{\cap_n^*} + \lambda_n \sum(\hat{\beta}_{\cap_n^*}^{*(n)})\right)^{-1}.$$

Si σ^2 est inconnue, on prend son estimateur dans le modèle complet. Pour les variables où $\hat{\beta}_j^{*(n)}$=0, les erreurs standards sont nulles (Tibshirani 1996[43] ; Fan et Li, 2001[12]).

3.2.4 Propriété d'oracle

Sous certaines conditions, Zou (2006)[51] a montré que l'estimateur du lasso adaptatif satisfait la propriéte d'oracle. En termes simples, la propriéte d'oracle signifie que l'estimateur du lasso adaptatif sélectionne les coefficients non nuls avec une probabilité tendant vers 1 et des composantes non nulles sont également estimées comme si le vrai modèle était connu, a priori (normalité asymptotique) (Fan et Li, 2001)[12].

Théorème 1 *Supposons que* $\lambda_n/\sqrt{N} \to 0$ *et* $\lambda_n N^{(\gamma-1)/2} \to \infty$. *Alors les estimations du lasso adaptatif satisfont les conditions :*
(1) convergence dans la sélection des variables : $lim_N P(\cap_n^* = \cap) = 1$,
(2) normalité asymptotique : $\sqrt{N}(\hat{\beta}_{\cap}^{*(n)} - \beta_{\cap}^*) \to_d N(0, \sigma^2 \times C_{11}^{-1})$,

$\cap = \{1, 2, ..., p_0\}$. Sous l'hypothèse que :
(a) $y_i = x_i\beta^* + \epsilon_i$, où ϵ_1,...,ϵ_n sont iid de moyenne 0 et de variance σ^2.
(b) $\frac{1}{n}X^TX \to$ C, matrice définie positive où C=$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$, et C_{11} est une matrice de dimension $p_0 \times p_0$. On suppose également que $\hat{\beta}$ est un estimateur convergent de racine N de β^*.

3.3 Lasso adaptatif pour l'issue

La description de cette méthode est tirée de l'article de Shortreed et Ertefaie (2017)[40].
Le lasso adaptatif pour l'issue est spécifiquement conçu pour l'inférence causale. Cette méthode de sélection des covariables vise à produire un estimateur du score de propension (PS) non biaisé et statistiquement efficace.

3.3.1 Estimation

Définissons les indices suivants :
C : indices des covariables associées à l'issue et à l'exposition.
P : indices des covariables liées à l'issue et non à l'exposition.
I : indices des covariables liées à l'exposition et non à l'issue.
S : indices des covariables ni liées à l'exposition, ni à l'issue.

Idéalement, la méthode PS inclut toutes les variables confondantes (X_C) pour éviter les biais et toutes les prédictives (X_P) pour augmenter l'efficacité statistique ; en excluant les prédicteurs de l'exposition (X_I) et des variables superflues (X_S). Un défi important serait alors d'imposer une pénalité plus lourde sur les prédicteurs de l'exposition que sur ceux de l'issue.

On utilise la régression logistique pour le modèle PS :

$$logit\{e(X,\hat{\alpha})\} = logit\{P(A=1|X,\hat{\alpha})\} = \sum_{j\in C}\hat{\alpha}_j X_j + \sum_{j\in P}\hat{\alpha}_j X_j. \tag{3.3}$$

Les estimations du lasso adaptatif pour l'issue sont données par :

$$\hat{\alpha}(OAL) = \underset{\alpha}{arg\,min}\left[\sum_{i=1}^{N}\{-a_i(X_i^T\alpha) + log(1+e^{X_i^T\alpha})\} + \lambda_n\sum_{j=1}^{p}\hat{\omega}_j|\alpha_j|\right], \tag{3.4}$$

$\hat{\omega}_j = |\tilde{\beta}_j|^{-\gamma}$, $\gamma > 1$ et $(\tilde{\beta}, \tilde{\theta}) = \text{argmin}_{\beta,\theta}\ \ln(\beta, \theta\,;\, \text{Y, X, A})$. On utilise $\tilde{\beta}$ pour référer aux coefficients estimés de la relation entre les covariables et l'issue conditionnellement au traitement avec $\tilde{\theta}$ le coefficient estimé correspondant au traitement. Ces estimations prennent aussi en compte les corrélations entre les covariables et l'issue.

Comportement asymptotique :
Supposons que $\lambda_n/\sqrt{N} \to 0$ et $\lambda_n N^{\gamma/2-1} \to \infty$, pour $\gamma > 1$, alors l'estimateur du lasso adaptatif pour l'issue satisfait la propriété d'oracle vue précédemment, sous les conditions de faible régularité. Autrement dit, cette méthode permet d'éliminer les variables prédisant l'exposition mais pas l'issue et les variables parasites du modèle. En plus, la normalité asymptotique garantit que les estimateurs pénalisés correspondant aux facteurs de confusion et aux indicateurs de l'issue ont un comportement similaire à celui des estimateurs du maximum de vraisemblance lorsque le vrai modèle est connu a priori. Soit $\text{M} = C \cup P$, les covariables incluses dans le modèle PS et $\text{M}^c = I \cup S$, les covariables à exclure. On assume sans perte de généralité l'ordre suivant pour les indices : $\text{M} = \{j : j \in C \cup P\} = \{1, 2, ..., d_0\}$ avec $d_0 < d = |C| + |P| + |I| + |S|$. $I(\alpha^*) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$ est la matrice d'information de Fisher, et I_{11} est la matrice d'information de

Fisher de dimension $d_0 \times d_0$ du modèle PS parcimonieux. (1) $\lim_N P\{\hat{\alpha}_j(OAL) = 0 | j \in I \cup S\}=1$, et (2) $\sqrt{N}\{\hat{\alpha}(OAL) - \alpha_M^*\} \to_d N(0, I_{11}^{-1})$.

3.3.2 Choix de λ_n

L'estimateur IPTW obtenu avec ATE, utilise le score de propension pour équilibrer la distribution des covariables entre les groupes d'exposition. La valeur du paramètre de réglage, λ_n est obtenue en minimisant la différence de moyenne des poids pondérés (wAMD) entre les groupes d'exposition :

$$wAMD(\lambda_n) = \sum_{j=1}^{d} |\tilde{\beta}_j| \left| \frac{\sum_{i=1}^n \hat{\tau}_i^{\lambda_n} X_{ij} A_i}{\sum_{i=1}^n \hat{\tau}_i^{\lambda_n} A_i} - \frac{\sum_{i=1}^n \hat{\tau}_i^{\lambda_n} X_{ij}(1 - A_i)}{\sum_{i=1}^n \hat{\tau}_i^{\lambda_n}(1 - A_i)} \right|,$$

$$\hat{\tau}_i^{\lambda_n} = \frac{A_i}{\hat{e}_i^{\lambda_n}\{X_i, \hat{\alpha}(OAL)\}} + \frac{1 - A_i}{1 - \hat{e}_i^{\lambda_n}\{X_i, \hat{\alpha}(OAL)\}}, \tag{3.5}$$

$\hat{\tau}^{\lambda_n}$ sont les IPTWs construits en ajustant un modèle PS dans lequel les variables sont choisies avec la méthode du lasso adaptatif pour l'issue, dénoté $\hat{\pi}^{\lambda_n}(.)$.

Jusqu'ici, les auteurs ont proposé la méthode du boostrap lissé pour l'estimation de l'erreur standard (Efron, 2014)[11]. Cette dernière est intensive du point de vue des calculs pour obtenir des inférences pour ce type de lasso[B].

3.4 Lasso bayésien

La description de cette méthode est essentiellement tirée des articles de Park et Casella (2008)[31] et Leng, Tran et Nott (2013)[25].

3.4.1 Présentation

Tibshirani (1996)[43] a remarqué que l'estimateur du lasso peut être considéré comme le mode de la distribution a posteriori de β, $\hat{\beta}_L = argmax_\beta p(\beta | y, \sigma^2, \tau)$, quand les p coefficients de régression ont une distribution a priori indépendante et identique de Laplace (i.e., double exponentielle),

$$p(\beta|\tau) = (\tau/2)^p exp(-\tau||\beta||_1), \tag{3.6}$$

et qu'une loi normale multivariée est prise pour la composante de la vraisemblance,

$$p(y|\beta, \sigma^2) = N(y|X\beta, \sigma^2 I_n).$$

B. *Le lien suivant permet d'accéder au code R implantant le lasso adaptatif pour l'issue :* https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12679.

Pour toutes valeurs fixées $\sigma^2 > 0$ et $\tau > 0$, le mode a posteriori de β est l'estimateur du lasso avec une pénalité $\lambda = 2\tau\sigma^2$.

Park et Casella (2008)[31] sont les précurseurs de la régression bayésienne du lasso. Les travaux antérieurs de Fernández et Steel (2010)[13] considèrent la loi a priori (3.6) comme un cas spécial dans la régression bayésienne générale, sans établir de liens avec la procédure du lasso. Toutes les deux approches utilisent un mélange d'échelle de représentations normales d'une distribution double exponentielle pour créer une formulation hiérarchique du modèle en introduisant un vecteur de variables d'échelle latentes θ. Finalement, on obtient des échantillons à partir de la distribution a posteriori jointe $p(\beta, \theta|\mathrm{y}, \sigma^2, \tau)$ via un échantillonneur de Gibbs (Tanner et Wong, 1987)[42] qui échantillonne itérativement à partir des distributions conditionnelles complètes $p(\beta|\theta, \mathrm{y}, \sigma^2, \tau)$ et $p(\theta|\beta, \mathrm{y}, \sigma^2, \tau)$. Park et Casella (2008)[31] ont étendu la régression bayésienne du lasso en introduisant des distributions a priori sur σ^2 et τ^2 pour prendre en compte l'incertitude dans ces hyper-paramètres.

Considérons le modèle hiérarchique suivant :

$$\begin{aligned} \mathbf{y}|\mu, \mathrm{X}, \beta, \sigma^2 &\sim N_n(\mu\mathbf{1}_n + \mathrm{X}\beta, \sigma^2 I_n) \\ \beta|\sigma^2, \tau_1^2, ..., \tau_p^2 &\sim N_p(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau) \qquad (3.7) \\ \mathbf{D}\tau &= diag(\tau_1^2, ..., \tau_p^2), \end{aligned}$$

$$\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2)d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} d\tau_j^2 \qquad (3.8)$$

Le schéma pour générer des échantillons avec l'échantillonneur de Gibbs à partir du modèle hiérarchique défini précédemment est le suivant. On génére une distribution conditionnelle complète de β suivant une loi normale multivariée de moyenne $\mathrm{A}^{-1}X^{'}\mathrm{y}$ et de matrice de covariance $\sigma^2 A^{-1}$, où $\mathrm{A}^{-1} = X^{'}\mathrm{X} + \mathrm{D}_\tau^{-1}$. La distribution conditionnelle complète de σ^2 est inverse gamma de paramètre de forme (n-1)/2 + p/2 et de paramètre d'intensité $(\mathrm{y}\text{-}\mathrm{X}\beta)^{'}(\mathrm{y}\text{-}\mathrm{X}\beta)/2 + \beta^{'} D_\tau^{-1}\beta/2$. Enfin, la distribution conditionnelle complète de $1/\tau_j^2$ est inverse gaussienne de moyenne $\tilde{\mu}_j = \lambda\sigma/|\beta_j|$ et de paramètre de forme $\tilde{\lambda} = \lambda^2$, où la densité inverse gaussienne est donnée par :

$$f(x) = \sqrt{\frac{\tilde{\lambda}}{2\pi}} x^{-3/2} exp\left\{-\frac{\tilde{\lambda}(x-\tilde{\mu}_j)^2}{2(\tilde{\mu}_j)^2 x}\right\}, \; x > 0.$$

On admet sans perte de généralité que y et X sont centrés.

3.4.2 Choix du paramètre du lasso bayésien

Deux approches peuvent être utilisées pour le choix du paramètre du lasso bayésien : la méthode du Bayes empirique (EB) et l'approche bayésienne hiérarchique. L'approche EB vise à estimer le paramètre λ via le maximum de vraisemblance marginal, alors que l'approche bayésienne hiérarchique utilise une loi a priori sur λ^2 permettant de faire des inférences sur les paramètres de réglages.

Nous nous intéressons ici à la seconde approche et considérons la classe des lois a priori gamma sur λ^2 de la forme :

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \lambda^2 > 0 \ (r > 0, \ \delta > 0). \tag{3.9}$$

L'avantage d'utiliser une telle loi a priori est de faciliter l'implantation de l'algorithme de l'échantillonneur de Gibbs. Il est aussi important que r et δ prennent des valeurs non nulles parce que dans le cas contraire, la loi a priori pour λ^2 serait $1/\lambda^2$ qui donne une loi a posteriroi impropre. La loi a priori (3.9) définit une distribution conditionnelle complète gamma pour λ^2, de paramètre de forme p+r et de paramètre d'intensité $\Sigma_{j=1}^{p}\tau_j^2/2+\delta$. Avec une telle spécification, λ^2 peut simplement joindre les autres paramètres dans l'échantillonneur de Gibbs car les distributions conditionnelles complètes des autres paramètres ne changent pas. La densité a priori de λ^2 devrait s'approcher de 0 suffisamment vite lorsque $\lambda^2 \to \infty$ mais devrait aussi être relativement plate et placer des probabilités élevées près de l'estimateur du maximum de vraisemblance.

Bien que complexe en calculs, le lasso bayésien fournit automatiquement des estimations d'intervalles pour tous les paramètres y compris la variance de l'erreur. Toutefois, en estimant que des coefficients non nuls, il perd la propriété attrayante de sélection de variables (Park et Casella, 2008)[31].

3.4.3 Exemple 4 (Lasso bayésien)

```
require(MASS)
require(MCMCpack)
require(SuppDists)
On déclare la taille de l'échantillon (n) et le nombre d'itération (niter)
et on initialise nos paramètres.
n = 1000
niter = 2000
beta = matrix(NA, nrow = niter, ncol = 2)
sigma = numeric(niter)
```

```
lambda = numeric(niter)
tau = matrix(NA, nrow = niter, ncol = 2) #Contient les tau^2
p = 2
set.seed(471984971)
X1 = rnorm(n)
X2 = rnorm(n)
Y = X1 + rnorm(n)
X = cbind(X1, X2)
Les valeurs initiales de sigma et tau sont définies en ajustant un modèle de
régression linéaire simple; la valeur initiale de lambda est choisie
arbitrairement, et les valeurs de delta de r sont fixées à 0.1.
p correspond au nombre de covariables.
mod = summary(lm(Y ~ -1 + X1 + X2))
beta[1,] = coef(mod)[c(2,3),1]
sigma[1] = mod$sigma^2
lambda[1] = 1
tau[1,] = diag(solve(t(X)%*%X))
delta = 0.1
r = 0.1
Xp = t(X)
XpX = t(X)%*%X
On met en exécution le schéma de l'échantillonneur Gibbs décrit plus haut
à travers cette boucle.
for(i in 2:niter){
 A = XpX + solve(diag(tau[i-1,]))
 beta[i,]=mvrnorm(n=1, mu=solve(A)%*%Xp%*%Y, Sigma=sigma[i-1]*solve(A))
 sigma[i]=rinvgamma(n=1, shape=(n-1)/2 + p/2, scale=t(Y - X%*%beta[i,])%*%
  (Y - X%*%beta[i,])/2 + t(beta[i,])%*%solve(diag(tau[i-1,]))%*%beta[i,]/2)
 tau[i,]=1/rinvGauss(n=p, nu=lambda[i-1]*sqrt(sigma[i])/abs(beta[i,]),
                        lambda=lambda[i-1]^2)
  lambda[i]=sqrt(rgamma(n=1, shape=p + r, rate=sum(tau[i,]/2) + delta))
}
*** Courbes de distribution
par(mfrow = c(2,1))
plot(beta[,1], type = "l")
plot(beta[,2], type = "l")
Il faut éliminer quelques itérations initiales, disons 100, mais 5 ou 10
serait probablement assez pour obtenir des distributions stationnaires.
Autrement dit, en effaçant les 100 premières valeurs simulées, on devrait
```

```
avoir un échantillon de la loi a posteriori désirée.
*** Courbes d'auto-corrélation
acf(beta[,1])
acf(beta[,2])
Il ne semble pas y avoir d'auto-corrélation, pas besoin
d'appliquer "d'amincissement" (thinning).
*** Moyennes a posteriori
mean(beta[101:2000,1])
mean(beta[101:2000,2])
*** Intervalles de crédibilité à 95%
quantile(beta[101:2000, 1], c(0.025, 0.975))
quantile(beta[101:2000, 2], c(0.025, 0.975))
*** Erreurs standard
sd(beta[101:2000,1])
sd(beta[101:2000,2])
```

NB : Avant d'exécuter la commande « require », assurez-vous d'abord d'avoir installé le package en question.

3.5 Lasso adaptatif bayésien

Cette méthode du lasso est décrite en se basant sur l'article de Leng et al., 2013 [25].
Avec le lasso adaptatif, différentes pénalités sont appliquées sur les différents coefficients. Zou (2006) [51] et Wang et al. (2007) [47] ont proposé des paramètres de pénalité de la forme $\lambda_j = \lambda\omega_j$, où les poids ω_j sont calculés à partir d'estimations préliminaires et le paramètre λ est choisi par validation croisée.

Dans la perspective bayésienne hiérarchique, aucune forme particulière n'est imposée sur les paramètres λ_j. Ces paramètres sont alors traités comme des variables aléatoires qui sont intégrées dans l'échantillonneur de Gibbs. Nous pouvons ainsi résumer les échantillons obtenus des paramètres λ_j à partir de leur distribution a posteriori. Ce qui semble moins demandant du point de vue computationnel que de calculer la valeur de λ avec le lasso adaptatif par validation croisée. Nous devons alors remplacer (3.8) dans la structure hiérarchique par une pénalité plus adaptative,

$$\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2) \prod_{j=1}^{p} \frac{\lambda_j^2}{2} e^{-\lambda_j^2 \tau_j^2/2}. \tag{3.10}$$

Intuitivement, l'estimateur du lasso en tant que mode a posteriori sera beaucoup plus précis si de faibles pénalités sont appliquées sur les coefficients les plus significatifs et de fortes pénalités

sont mises sur les coefficients les moins significatifs. En intégrant sur les τ_j^2s dans (3.7) et (3.10), la distribution a priori conditionnelle de β sachant σ^2 est :

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda_j}{2\sqrt{\sigma^2}} e^{-\lambda_j|\beta_j|/\sqrt{\sigma^2}}.$$

De manière similaire que Park et Casella (2008)[31], la loi a posteriori $\pi(\beta, \sigma^2|y)$, pour tout λ_js est unimodale. L'unimodalité est importante car cela amène l'échantillonneur de Gibbs à converger plus rapidement et rend plus significative l'estimation des paramètres.

3.5.1 Choix des paramètres du lasso adaptatif bayésien

Tout comme pour le lasso bayésien, les paramètres λ_j pour le lasso bayésien adaptatif sont choisis à travers les deux approches que sont la méthode du Bayes empirique et l'approche bayésienne hiérarchique. Nous considérons également une loi a priori gamma pour λ_j^2 comme suit :

$$\pi(\lambda_j^2) = \frac{\delta^r}{\Gamma(r)} (\lambda_j^2)^{r-1} e^{-\delta\lambda_j^2}. \tag{3.11}$$

Avec une telle loi a priori, λ_j^2 suit une distribution conditionnelle complète gamma de paramètre de forme 1+r et de paramètre d'intensité $\tau_j^2 + \delta$. Cette spécification permet à λ_j^2 de joindre les autres paramètres dans l'échantillonneur de Gibbs.

Leng et al., 2013[25] considèrent comme premier choix d'utiliser des valeurs petites pour les hyper-paramètres r et δ pour avoir des lois *a priori* non informatives. Alternativement, r peut être fixé et δ estimé en utilisant une approche Bayes empirique comme défini par Casella (2001)[7] :

$$\delta^{(k)} = \frac{pr}{\Sigma_{j=1}^{p} E_{\delta(k-1)}(\lambda_j^2|y)}.$$

3.5.2 Exemple 5 (Lasso adaptatif bayésien)

```
require(MASS)
require(MCMCpack)
require(SuppDists)
n = 1000
niter = 2000
beta = matrix(NA, nrow = niter, ncol = 2)
sigma = numeric(niter)
tau = matrix(NA, nrow = niter, ncol = 2)
lambda = matrix(NA, nrow = niter, ncol = 2)
delta = 0.1
```

```
r = 0.1
p = 2
set.seed(471984971)
X1 = rnorm(n)
X2 = rnorm(n)
Y = X1 + rnorm(n)
X = cbind(X1, X2)
*** Valeur de départ
mod = summary(lm(Y ~ -1 + X1 + X2))
beta[1,] = coef(mod)[c(2,3),1]
sigma[1] = mod$sigma^2
tau[1,] = diag(solve(t(X)%*%X))
lambda[1,] = 1
Xp = t(X)
XpX = t(X)%*%X
for(i in 2:niter){
  A = XpX + solve(diag(tau[i-1,]))
beta[i,]=mvrnorm(n=1, mu=solve(A)%*%Xp%*%Y, Sigma=sigma[i-1]*solve(A))
sigma[i]=rinvgamma(n=1, shape=(n-1)/2 + p/2, scale=t(Y - X%*%beta[i,])%*%
      (Y - X%*%beta[i,])/2 + t(beta[i,])%*%solve(diag(tau[i-1,]))%*%beta[i,]/2)
tau[i,]=pmin(1/rinvGauss(n=p, nu=lambda[i-1,]*sqrt(sigma[i])/abs(beta[i,]),
                         lambda=lambda[i-1,]^2),10000)
lambda[i,]=rgamma(n=p, shape=1 + r, rate=tau[i,] + delta)^0.5
par(mfrow = c(2,1))
plot(beta[,1], type = "l")
plot(beta[,2], type = "l")
acf(beta[,1])
acf(beta[,2])
*** Moyennes a posteriori
mean(beta[101:2000,1])
mean(beta[101:2000,2])
*** Intervalles de crédibilité à 95%
quantile(beta[101:2000, 1], c(0.025, 0.975))
quantile(beta[101:2000, 2], c(0.025, 0.975))
}
```

Chapitre 4

Estimation bayésienne du lasso adaptatif pour l'issue

Nous allons d'abord introduire dans ce chapitre le lasso bayésien adaptatif pour le modèle linéaire généralisé. Cette généralisation vise à construire un modèle du score de propension qui nous permet par la suite d'estimer l'effet du traitement moyen par l'IPTW. Ensuite, nous aborderons l'estimation bayésienne du lasso adaptatif pour l'issue, pour finir avec une application des différentes méthodes de lasso aux données de l'étude PREDISE.

4.1 Lasso adaptatif bayésien du modèle linéaire généralisé

La description de cette méthode est tirée de l'article de Leng et al., 2013 [25].
Wang et Leng (2007) [47] utilise une approximation des moindres carrés de la log-vraisemblance (-L(β)) dans le cadre de la généralisation du lasso bayésien adaptatif :

$$L(\beta) \approx L(\tilde{\beta}) + \frac{\partial L(\tilde{\beta})}{\partial \beta}(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})' \frac{\partial^2 L(\tilde{\beta})}{\partial \beta \partial \beta'}(\beta - \tilde{\beta})$$

$$= constante \ + \frac{1}{2}(\beta - \tilde{\beta})' \hat{\Sigma}^{-1}(\beta - \tilde{\beta}),$$

où $\tilde{\beta}$ est l'estimation du maximum de vraisemblance de β et $\hat{\Sigma}^{-1} := \partial^2 L(\tilde{\beta}) / \partial \beta^2$. Avec le lasso bayésien adaptatif pour le modèle linéaire généralisé, la distribution conditionnelle de y peut être approximée par :

$$y|\beta \sim \exp(-\tfrac{1}{2}(\beta - \tilde{\beta})' \hat{\Sigma}^{-1}(\beta - \tilde{\beta})).$$

Ainsi, dans le modèle hiérarchique, seule la distribution de y sera mise à jour comme on peut le voir dans la formulation hiérarchique suivante

$$\mathrm{y}|\beta \sim \exp(-\tfrac{1}{2}(\beta - \tilde{\beta})' \hat{\Sigma}^{-1}(\beta - \tilde{\beta})),$$

$$\beta|\tau^2 \sim N_p(0, D_\tau),\ D_\tau = diag(\tau^2),$$

$$\tau^2|\lambda^2 \sim \prod_{j=1}^{p} \frac{\lambda_j^2}{2} e^{-\lambda_j^2 \tau_j^2/2},$$

$$\lambda^2 \sim \prod_{j=1}^{p} (\lambda_j^2)^{r-1} e^{-\delta\lambda_j^2},$$

où $\tau^2 := (\tau_1^2, ..., \tau_p^2)'$, $\lambda^2 := (\lambda_1^2, ..., \lambda_p^2)'$. Par contre, la distribution de σ^2 ne figure plus dans la hiérarchie. Les distributions conditionnelles sont spécifiées par :

$$\beta|y, \tau^2, \lambda^2 \sim N_p((\hat{\Sigma}^{-1} + D_\tau^{-1})^{-1}\hat{\Sigma}^{-1}\tilde{\beta},\ (\hat{\Sigma}^{-1} + D_\tau^{-1})^{-1}),$$

$$\frac{1}{\tau_j^2} = \gamma j|y, \beta, \lambda^2 \sim \text{inverse-Gaussian}(\frac{\lambda_j}{|\beta_j|}, \lambda_j^2), \text{ j=1,...,p},$$

$$\lambda_j^2|y, \beta, \tau^2 \sim \text{gamma}(\text{r+1}, \delta + \frac{\tau_j^2}{2}), \text{ j=1,...,p.}$$

4.1.1 Exemple 6 (Lasso adaptatif bayésien du modèle linéaire généralisé)

```
set.seed(282323)
n = 10000
X1 = rnorm(n)
Y = rbinom(n, 1, plogis(-0.2+0.5*X1))
X2 = rnorm(n)
require(MASS)
require(MCMCpack)
require(SuppDists)
niter = 2000
beta = matrix(NA, nrow = niter, ncol = 2)
tau = matrix(NA, nrow = niter, ncol = 2)
lambda = matrix(NA, nrow = niter, ncol = 2)
delta = 0.1
r = 0.1
p = 2
X = cbind(X1, X2)
***Valeurs de départ
ps = summary(glm(Y ~ -1 + X1 + X2, family=binomial(link=logit)))
beta[1,] = coef(ps)[c(2,3),1]
tau[1,] = diag(solve(t(X)%*%X))
lambda[1,] = 1
```

```
sigma.Chapeau <- solve(ps$cov.unscaled[c(2,3),c(2,3)]) # Contient l'inverse
                                        #de la matrice de covariance.
for (i in 2:niter){
 beta[i,]=mvrnorm(n=1, mu=solve(sigma.Chapeau + solve(diag(tau[i-1,])))%*%
 sigma.Chapeau%*%beta[1,], Sigma=solve(sigma.Chapeau + solve(diag(tau[i-1,]))))
 tau[i,]=1/rinvGauss(n=p, nu=lambda[i-1,]/abs(beta[i,]), lambda=lambda[i-1,]^2)
 lambda[i,]=rgamma(n=p, shape=r + 1, rate=delta + tau[i,]/2)^0.5
}
par(mfrow = c(2,1))
plot(beta[,1], type = "l")
plot(beta[,2], type = "l")
***Courbes d'auto-corrélation
acf(beta[,1])
acf(beta[,2])
***Moyennes a posteriori
mean(beta[101:2000,1])
mean(beta[101:2000,2])
***Intervalles de crédibilite à 95%
quantile(beta[101:2000, 1], c(0.025, 0.975))
quantile(beta[101:2000, 2], c(0.025, 0.975))
```

4.2 IPTW bayésienne

Tel que souligné dans le chapitre 2, l'IPTW permet d'avoir une certaine population dans laquelle les déséquilibres sur les covariables mesurées entre les groupes de traitement sont éliminés. Une version bayésienne d'une telle procédure peut être liée à l'échantillonnage dans cette population cible en utilisant la vraisemblance pondérée pertinente (relevance-weighted likelihood, REWL) de Hu et Zidek (2002)[19] ou de Wang (2006)[48], ou le bootstrap de la vraisemblance pondérée de Newton & Raftery, 1994[30] (Saarela et al., 2015)[37]. Cette méthode devrait donc nous permettre au final de simuler des échantillons à partir de la population cible avec les données observées.

Soit f_i une fonction de densité de probabilité inconnue sur des données Y_i, avec $Y_i \perp\!\!\!\perp Y_j$, i≠j. La vraisemblance pondérée pertinente est alors : $REWL(\theta)=\prod_{i=1}^{n} p(y_i;\theta)^{\lambda_i}$, où $\lambda_1,\ldots,\lambda_n$ sont les poids de la vraisemblance (Wang, 2006)[48]. Cette dernière est particulièrement utile lorsqu'un échantillon de la population cible n'est pas disponible, mais que des échantillons d'autres populations sont pertinents pour en apprendre davantage sur la population cible

(Saarela et al., 2015)[37].

Dans le cas qui nous intéresse, on cherche à estimer le paramètre θ d'un modèle structurel marginal reliant les issues contrefactuelles à l'exposition, $p(y_i^a|\theta)$. L'idée de l'échantillonnage d'importance dans ce cas est d'estimer ce paramètre θ en utilisant un modèle analogue au précédent, mais modélisant la réponse observée $p(y|a,\theta)$ où les données sont pondérées de façon à simuler la population cible où il y a une distribution équilibrée des covariables entre les groupes d'exposition. Ainsi, sous les conditions de positivité et d'échangeabilité conditionnelle, on peut faire cette estimation de θ en maximisant la fonction pseudo-vraisemblance pondérée suivante :

$$q(\theta; v, \gamma, \alpha) \equiv \prod_{i=1}^{n} p(y_i \mid A_i, \theta)^{\omega_i}, \tag{4.1}$$

où $\omega_i = \frac{p(A_i \mid \alpha)}{p(A_i \mid X_i, \gamma)}$ définissent les poids "stabilisés" ; v=(X,Y,A) et, α et γ représentent les coefficients d'une régression logistique marginale et conditionnelle de A. Dans le cas où les poids ω_i sont fixés, on peut représenter un échantillon aléatoire des n sujets observés avec une contribution égale à l'information à travers cette fonction de vraisemblance (Saarela et al., 2015)[37]

$$q(\theta; v, \gamma, \alpha, \pi) = \prod_{i=1}^{n} p(y_i \mid A_i, \theta)^{n\pi_i\omega_i}, \tag{4.2}$$

où $\pi \equiv (\pi_1,\ldots,\pi_n) \sim \text{Dirichlet}(1, \ldots, 1)$.

Les poids Dirichlet sont très convenables à manipuler comme ils ne sont jamais égaux à zéro. En plus, ils forment une base naturelle pour la simulation bayésienne grâce à leurs propriétés pour des données multinomiales (Newton & Raftery, 1994)[30].

Dans le cas où les vraies valeurs de α et γ ne sont pas connues, c'est à dire que les poids ω_i ne sont plus fixés, alors Saarela et al., 2015[37] ont démontré l'expression suivante pour les poids grâce aux propriétés bayésiennes :

$$\begin{aligned} \omega_i^* &= \frac{\int_\alpha p(A_i^* \mid \alpha, O)\, p(\alpha \mid A, O) d\alpha}{\int_\gamma p(A_i^* \mid X_i^*, \gamma, O)\, p(\gamma \mid X, A, O) d\gamma} \\ &= \frac{E_\alpha\Big[p(A_i^* \mid \alpha, O) \mid A, O\Big]}{E_\gamma\Big[p(A_i^* \mid X_i^*, \gamma, O) \mid X, A, O\Big]}. \end{aligned} \tag{4.3}$$

La notation O permet de représenter le mécanisme observationnel de génération des données, et $v_i^* = (X_i^*, Y_i^*, A_i^*)$ désignent des observations prédites sous le régime expérimental E. L'expression simplifiée dans (4.3) suppose que les facteurs confondants potentiels non observés

sont indépendants de l'exposition. Les poids w_i^* correspondent ainsi au rapport de l'espérance a posteriori des prédictions d'observations futures A_i^* selon le modèle $p(A|\alpha)$, sur la même espérance a posteriori selon le modèle $p(A|X, \gamma)$. En outre, une version analogue à la fonction de vraisemblance présentée dans l'équation 4.2 devient

$$E\left[l(y_i^* \mid A_i^*, \theta) \mid v, E\right] = \sum_{i=1}^{n} \pi_i \omega_i^* l(y_i \mid A_i, \theta), \tag{4.4}$$

où $l(y_i^* \mid A_i^*, \theta) \equiv log\, p(y_i^* \mid A_i^*, \theta)$. Alors, l'estimateur du maximum de vraisemblance pondérée de θ est $\hat{\theta}(v; \pi) \equiv arg \underset{\theta}{max}\left[\sum_{i=1}^{n} \pi_i \omega_i^* l(y_i \mid A_i, \theta)\right]$.

Autrement dit, on peut obtenir un échantillon a posteriori de θ en suivant la procédure suivante : 1) effectuer une estimation bayésienne de $p(A_i|\alpha)$ et $p(A_i|X_i, \gamma)$, 2) calculer les poids w_i^* à partir des moyennes a posteriori des modèles obtenus à l'étape 1, 3) prendre des échantillons aléatoires de π dans une loi Dirichlet et calculer la valeur $\hat{\theta}$ qui maximise la vraisemblance pondérée (4.4), 4) effectuer les inférences a posteriori sur θ à partir des échantillons obtenus à l'étape 3.

4.3 Algorithme proposé pour l'estimation bayésienne du lasso adaptatif pour l'issue

La version bayésienne du lasso adaptatif pour l'issue que nous proposons consiste d'abord à effectuer une modélisation bayésienne du score de propension à l'aide d'une version modifiée du lasso adaptatif bayésien généralisé présenté à la section 4.1. Le résultat de cette modélisation est ensuite utilisé pour construire le dénominateur des poids dans l'approche IPTW bayésienne que nous venons de présenter. Cette version bayésienne du lasso adaptatif pour l'issue est également inspirée dans sa structure de la version fréquentiste du lasso adapatatif pour l'issue.

Concrètement, nous proposons d'utiliser une distribution a priori informative pour les paramètres de régularisation λ_j du lasso adaptatif bayésien généralisé. Cette distribution est de nature bayes empirique, puisqu'elle utilise des informations provenant des données. Le paramètre λ_j détermine le niveau de régularisation du coefficient associé à X_j dans le modèle du score de propension. Plus sa valeur est élevée, plus le coefficient est contracté vers 0. Au contraire, une valeur de λ_j proche de 0 indique que le coefficient n'est pas contracté.

Nous proposons donc de poser $\lambda_j^2 \sim \Gamma(r_j, \delta_j)$ avec $r_j = \left[n^d |\hat{\tilde{\beta}}_j|^{-\gamma}\right]^2$ et $\delta_j = n\sqrt{r_j}$, où $\tilde{\beta}_j$ est la vraie valeur du coefficient associé à X_j dans le modèle de Y en fonction de A et X, et $\hat{\tilde{\beta}}_j$ est sa

valeur estimée. Nous supposons que l'estimateur $\hat{\tilde{\beta}}_j$ est convergent avec taux de convergence $\sqrt{n}$. L'espérance et la variance a priori de λ_j^2 sont données par :

$$\mathbb{E}(\lambda_j^2) = \frac{r_j}{\delta_j} = \frac{\sqrt{r_j}}{n} = n^{d-1}|\hat{\tilde{\beta}}_j|^{-\gamma}$$
$$Var(\lambda_j^2) = \frac{r_j}{\delta_j^2} = \frac{1}{n^2}$$

Notons que la forme de l'espérance a priori est similaire à celle de la pénalisation de α_j dans le lasso adaptatif pour l'issue fréquentiste, c'est-à-dire $\lambda_n|\hat{\tilde{\beta}}_j|^{-\gamma}$, où les auteurs proposent de prendre $\lambda_n = n^d$.

Nous allons maintenant montrer que ce choix de distribution a priori pour λ_j a les propriétés suivantes lorsque $n \to \infty$ pour des valeurs adéquatement choisies de d et γ :

$$\mathbb{E}(\lambda_j^2) \to 0 \text{ si } \tilde{\beta}_j \neq 0,$$
$$\mathbb{E}(\lambda_j^2) \to \infty \text{ si } \tilde{\beta}_j = 0,$$
$$Var(\lambda_j^2) \to 0.$$

Ainsi, la distribution a priori devient de plus en plus informative lorsque n augmente, car la variance a priori diminue. Par ailleurs, si la variable j n'est pas associée à Y, la distribution a priori a une pénalité qui tend vers l'infini. Par contre, si la variable j est associée à Y, la distribution a priori a une pénalité qui tend vers 0. Ainsi, cette distribution a priori force l'inclusion des prédicteurs purs de Y ainsi que des variables confondantes (X_P et X_C), tel que désiré.

Preuve
Comme mentionné plus haut, $\mathbb{E}(\lambda_j^2) = n^{d-1}|\hat{\tilde{\beta}}_j|^{-\gamma}$. Si $\tilde{\beta}_j \neq 0$, $n^{d-1}|\hat{\tilde{\beta}}_j|^{-\gamma} = O(n^{d-1})$, où O est la notation "grand-O" usuelle, c'est-à-dire que $n^{d-1}|\hat{\tilde{\beta}}_j|^{-\gamma}$ se comporte asymptotiquement de la même façon que n^{d-1}. Lorsque $d < 1$, on obtient donc $\lim_{n\to\infty} \mathbb{E}(\lambda_j^2) = 0$.

Si $\tilde{\beta}_j = 0$, alors $|\hat{\tilde{\beta}}_j| = O(n^{-1/2})$. Ainsi, $n^{d-1}|\hat{\tilde{\beta}}_j|^{-\gamma} = O(n^{d-1-\gamma/2})$. On obtient donc $\lim_{n\to\infty} \mathbb{E}(\lambda_j^2) = \infty$ lorsque $d > \gamma/2 + 1$.

Finalement, puisque $Var(\lambda_j^2) = n^{-2}$, il est évident que $\lim_{n\to\infty} Var(\lambda_j^2) = 0$.■

Nous proposons en particulier de choisir $d = -1$ et $\gamma = -8$, ce qui fait en sorte que $\mathbb{E}(\lambda_j^2) = O(n^{-2})$ lorsque $\tilde{\beta}_j \neq 0$ et $\mathbb{E}(\lambda_j^2) = O(n^2)$ lorsque $\tilde{\beta}_j = 0$.

Dans l'exemple ci-dessous portant sur des données simulées, l'algorithme proposé permet d'identifier le modèle ciblé, soit celui incluant uniquement la variable confondante et la variable prédictrice de l'issue.

4.3.1 Exemple 7 (Lasso adaptatif pour l'issue bayésien)

```
require(MASS)
require(MCMCpack)
require(SuppDists)
options(scipen = 999)

expit = plogis
n = 10000
XS = rnorm(n)
XC = rnorm(n)
XI = rnorm(n)
XP = rnorm(n)
A = rbinom(n, 1, p = expit(-0.2 + XI + 0.2*XC))
Y = XC + XP + A + rnorm(n)
X = cbind(1, XC, XP, XS, XI)

niter = 20000
n.update = 1000
p.update = 10000
p = 5
lambda = matrix(NA, nrow = niter, ncol = p)
beta = matrix(NA, nrow = niter, ncol = p)
sigma = numeric(niter)
tau = matrix(NA, nrow = niter, ncol = p)
w = numeric(n)
theta = numeric(niter)

modA = glm(A ~ X - 1, family = "binomial")
modA.reduit = glm(A ~ 1 + XC + XP, family = "binomial")
modY <- summary(lm(Y ~ A + X - 1))
alpha = coef(modA)
Sigma.inv = solve(vcov(modA))
beta.temp = matrix(NA, nrow = n.update, ncol = p)
Sigma.temp = array(NA, dim = c(p, p, n.update))
```

```
Xnew = X

# Valeur de depart
beta[1,] = alpha
tau[1,] = rep(1, p)
lambda[1,] = rep(1, p)

r = (n**-1*abs(coef(modY)[2:(p+1),1])**-8)**2
r[1] = 0
delta = n*sqrt(r)

k = 0
for(i in 2:niter)
{
  Sigma = solve(Sigma.inv + diag(1/tau[i-1,]))
  mu = Sigma%*%Sigma.inv%*%alpha
  beta[i,] = mvrnorm(n = 1, mu = mu, Sigma = Sigma)
  tau[i,] = pmax(pmin(1/rinvGauss(n = p, nu = lambda[i-1,]/abs(beta[i,]),
                                  lambda = lambda[i-1,]**2), 10**16), 10**-16)
  lambda[i,] = sapply(1:p, FUN = function(x){rgamma(1, shape = 1 + r[x],
                                          rate = tau[i,x]/2 + delta[x])**0.5})
  k = k + 1
  beta.temp[k,] = mu
  if(k == n.update){
    index = abs(p.update*colMeans(beta.temp)) < abs(alpha)
    Xnew[,index] = Xnew[,index] + rnorm(n*sum(index))
    modA.new = glm(A~Xnew-1, family = "binomial")
    alpha = coef(modA.new)
    Sigma.inv = solve(vcov(modA.new))
    k = 0
    print(alpha)
  }
}

plot(beta[,1], type = "l")
plot(beta[,2], type = "l")
plot(beta[,3], type = "l")
plot(beta[,4], type = "l")
plot(beta[,5], type = "l")
```

```r
burn = 4000

# Coefficients + sd du lasso
cbind(colMeans(beta[burn:niter,]), apply(beta[burn:niter,], 2, sd))

# Vrais coefficients cibles
summary(modA.reduit)$coef[,1:2]

# Moyennes a posteriori de lambda
colMeans(lambda[burn:niter,])


alpha.star = colMeans(beta[burn:niter,])
pred2 = expit(Xnew%*%alpha.star)

mod.A0 = glm(A~1, family = binomial(link = "logit"))
pred = predict(mod.A0, type = "res")

w = A*pred/pred2 + (1 - A)*(1 - pred)/(1 - pred2)

for(i in 1:niter)
{
  Pi = as.numeric(rdirichlet(1, alpha = rep(1, n)))
  poids = Pi*w
  mod = lm(Y~A, weights = poids)
  theta[i] = coef(mod)[2]
}
mean(theta)
sd(theta)
quantile(theta, c(0.025, 0.975))
```

Avec le tableau ci-après, on constate que les vrais coefficients cibles sont très proches de ceux obtenus dans les résultats.

Toutefois, on a dû modifier l'algorithme pour mettre à jour les coefficients utilisés pour l'approximation linéaire.

Tableau 4.1 – Coefficients cibles versus estimés par notre méthode

	Coefficients cibles	Coefficients estimés	Ecarts
constante	-0.1467	-0.1374	-0.0093
XC	0.1384	0.1318	0.0066
XP	0.0234	0.0224	0.0010
XS	0.0000	0.0000	0.0000
XI	0.0000	0.0001	-0.0001

4.4 Illustration des différentes méthodes du lasso avec les données PREDISE

Il s'agira de présenter dans cette section les résultats obtenus en ajustant pour les différentes méthodes du lasso, un modèle ayant comme variable dépendante le logarithme de la pression sanguine systolique en mm Hg et en considérant la saine alimentation comme notre variable indépendante principale.

Tous les facteurs confondants potentiels sont à la base catégoriels sauf ces deux : « Nombre d'activité physique par semaine » et « Score de biais de désirabilité sociale (Paulhus, 1991) ». Nous avons donc dichotomisé les différentes catégories de chaque facteur confondant potentiel en choisissant une catégorie de référence de façon arbitraire.

Hormis le lasso adaptatif pour l'issue, les différentes méthodes donnent des résultats assez similaires en terme d'estimation du coefficient associé à la saine alimentation, mais aussi en terme d'erreur-type et d'intervalle de confiance. La saine alimentation est associée à une réduction de la tension artérielle lorsqu'on considère les résultats du lasso adaptatif pour l'issue qui sont statistiquement significatifs. Cette réduction est de 3.95% chez les personnes qui ont une bonne alimentation [A]. Les résultats des autres méthodes ne sont pas statistiquement significatifs (Tableau 4.2).

Le temps d'exécution des programmes des différentes méthodes a été aussi très variable. Ainsi, le lasso ordinaire a pris 1.568 secondes ; le lasso adaptatif 1.425 secondes ; le lasso adaptatif pour l'issue 2646.307 secondes, soit 44.105 minutes, le lasso bayésien a pris 31.691 secondes avec 20000 itérations ; le lasso adaptatif bayésien a pris 31.321 secondes avec 20000 itérations. Enfin le lasso adaptatif pour l'issue bayésien a pris 41.298 secondes avec 20000 itérations.

A. Avec la transformation log, l'exponentiel des coefficients estimés s'interprétent comme un rapport de moyenne géométrique. Pour plus de détails, voir https://www.bmj.com/content/312/7038/1079?ijkey=3ebafff648a1b4acad09ebde4259cf3d01dd3953&keytype2=tf_ipsecsha

Tableau 4.2 – Illustration des différentes méthodes du lasso avec les données PREDISE

Estimateur / Méthode	β C-HEI	Erreur-type	Intervalle de confiance à 95%
Lasso ordinaire	-0.0115	0.0074	-0.0260 ;0.0030
Lasso adaptatif	-0.0096	0.0064	-0.0221 ;0.0029
Lasso adaptatif pour l'issue	-0.0403	0.0085	-0.0570 ;-0.0236
Lasso bayésien	-0.0100	0.0071	-0.0244 ;0.0032
Lasso adaptatif bayésien	-0.0103	0.0073	-0.0246 ;0.0039
Lasso adaptatif pour l'issue bayésien	-0.0162	0.0085	-0.0328 ;0.0004

Le tableau 4.3 présente les covariables sélectionnées par les différentes méthodes. La sélection des covariables par le lasso ordinaire et le lasso adaptatif semble refléter les différences observées dans la distribution des covariables entre le groupe ayant une bonne alimentation et le groupe n'ayant pas une bonne alimentation (Tableau 1.2). Aussi, les méthodes bayésiennes du lasso qui ne sont pas des méthodes de sélection de variables n'ont exclu aucune covariable. Par ailleurs, le lasso adaptatif pour l'issue choisit uniquement la variable indépendante principale.

Tableau 4.3 – Les covariables choisies par les différentes méthodes

Covariables	Lasso ordinaire	Lasso adaptatif	Lasso adaptatif pour l'issue
C-HEI	x	x	x
Age : 35-49 ans	x	x	
Age : 50-65 ans	x	x	
Sexe : femme	x	x	
Niveau d'éducation : Cegep	x	x	
Niveau d'éducation : Université	x	x	
Revenu : 30000 à $<$ 60000$,CAD		x	
Revenu : 60000 à $<$ 90000$,CAD	x	x	
Revenu : 90000+		x	
Ethnicité : Africain-américain	x	x	
Ethnicité : Hispanique	x	x	
Ethnicité : Autre	x	x	
Centre de recruitement : ECOGENE-Ch	x	x	
Centre de recruitement : INAF		x	
Centre de recruitement : IRCM	x	x	
Centre de recruitement : UQTR	x	x	
Indice de masse corporelle : Surpoids(25-30)	x	x	
Indice de masse corporelle : Obèse($>$30)	x	x	
Statut fumeur : autrefois	x	x	
Statut fumeur : jamais		x	
Suivre un régime alimentaire particulier : Oui	x	x	
Consommation autodéclarée de suppléments alimentaires : Oui	x	x	
Compétences informatiques : Moyen			
Compétences informatiques : Fort			
Consommation autodéclarée de médicaments : Non	x	x	
Nombre d'activité physique par semaine	x	x	
Score de biais de désirabilité sociale (Paulhus, 1991)	x	x	

Conclusion

Le lasso adaptatif pour l'issue est une méthode efficiente de sélection de variables en inférence causale mais il est désavantagé par la lourdeur computationnelle de la procédure des inférences. Ce mémoire visait donc à proposer une nouvelle méthode de régression pour le lasso adaptatif pour l'issue basée sur les approches bayésiennes. Nos résultats vont dans le sens de valider notre hypothèse selon laquelle la nouvelle méthode que nous avons développée est moins exigeante du point de vue computationnel que le recours au boostrap lissé pour faire les inférences avec le lasso adaptatif pour l'issue. En effet, il a fallu seulement 41.298 secondes avec 20000 itérations pour produire une estimation de notre paramètre d'intérêt avec notre nouvelle méthode, alors que le lasso adaptatif pour l'issue a pris 44.105 minutes.

Le travail accompli dans ce mémoire constitue une première étape importante dans le développement d'une méthode bayésienne pour le lasso adaptatif pour l'issue. Les propriétés démontrées pour la loi a priori de λ_j^2 n'assurent cependant pas la propriété d'oracle, car les propriétés de la distribution a posteriori n'ont pas été établies. Il faudrait également trouver une solution plus élégante au problème de l'approximation linéaire LQA qui n'est pas très bonne lorsque certains coefficients sont très pénalisés. Par ailleurs, les résultats obtenus de nos application dépendent du choix d'échelle de Y et de X. Il est donc important de standardiser ces variables pour rendre les résultats insensibles au choix d'échelle.

Malgré qu'on ait illustré sur des données simulées que la méthode proposée permet d'atteindre les résultats attendus, il serait intéressant d'étudier plus en profondeur le comportement de cette méthode en fonction de différents facteurs comme par exemple la force des coefficients dans les deux modèles, de n et de p, de la corrélation entre les variables, à partir d'études de simulations. Enfin, on pense qu'un problème d'optimisation est à l'origine de l'exclusion des covariables par le lasso adaptatif pour l'issue.

Bibliographie

[1] An, W. (2010). 4. Bayesian propensity score estimators : Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1) :151–189.

[2] Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation*, 38(6) :1228–1234.

[3] Austin, P. C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29(20) :2137–2148.

[4] Austin, P. C. and Mamdani, M. M. (2006). A comparison of propensity score methods : a case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25(12) :2084–2106.

[5] Blanchet, C., Plante, C., and Rochette, L. (2009). La consommation alimentaire et les apports nutritionnels des adultes québécois. Technical report, Institut national de santé publique Québec.

[6] Brassard, D., Laramée, C., Corneau, L., Bégin, C., Bélanger, M., Bouchard, L., Couillard, C., Desroches, S., Houle, J., Langlois, M.-F., et al. (2018). Poor adherence to dietary guidelines among french-speaking adults in the province of Quebec, Canada : The predise study. *Canadian Journal of Cardiology*, 34(12) :1665–1673.

[7] Casella, G. (2001). Empirical Bayes gibbs sampling. *Biostatistics*, 2(4) :485–500.

[8] Diop, S. A. (2019). Comparing inverse probability of treatment weighting methods and optimal nonbipartite matching for estimating the causal effect of a multicategorical treatment.

[9] Duchesne, T. (2012). Stt 7140 statistique bayésienne. Recueil inédit, Université Laval.

[10] Duchesne, T. (2018). Stt 7120 théorie et applications des méthodes de régression. Recueil inédit, Université Laval.

[11] Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507) :991–1007.

[12] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456) :1348–1360.

[13] Fernandez, C. and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16(1) :80–101.

[14] Garriguet, D. (2009). Diet quality in canada. *Health Reports*, 20(3) :41.

[15] Hernan, M. and Robins, J. (2020). *Causal Inference : What If*. Boca Raton : Chapman & Hall/CRC.

[16] Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4) :1161–1189.

[17] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260) :663–685.

[18] Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, 52(3) :1413–1429.

[19] Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, 30(3) :347–371.

[20] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007) :453–461.

[21] Judkins, D. R., Morganstein, D., Zador, P., Piesse, A., Barrett, B., and Mukhopadhyay, P. (2007). Variable selection and raking in propensity scoring. *Statistics in Medicine*, 26(5) :1022–1033.

[22] Kaplan, D. and Chen, J. (2012). A two-step Bayesian approach for propensity score analysis : Simulations and case study. *Psychometrika*, 77(3) :581–609.

[23] Lawson, C. L. and Hanson, R. (1974). Linear least squares with linear inequality constraints solving least squares problems.

[24] Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 165(1) :59–82.

[25] Leng, C., Tran, M.-N., and Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2) :221–244.

[26] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects : a comparative study. *Statistics in Medicine*, 23(19) :2937–2960.

[27] McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics*, 6(2).

[28] McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28(1) :94–112.

[29] Meinshausen, N. and Buhlmann, P. (2004). Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.

[30] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society : Series B (Methodological)*, 56(1) :3–26.

[31] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482) :681–686.

[32] Robert, C. P. and Casella, G. (c2010). *Monte Carlo statistical methods.* Springer texts in statistics. Springer, New York, 2nd ed edition. Bibliogr. : p. [591]-622.

[33] Robins, J. M. and Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123(3) :392–402.

[34] Rolling, C. A. and Yang, Y. (2013). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(4) :749–769.

[35] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398) :387–394.

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

[37] Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics*, 71(2) :279–288.

[38] Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4) :512.

[39] Schwingshackl, L., Schwedhelm, C., Hoffmann, G., Knüppel, S., Iqbal, K., Andriolo, V., Bechthold, A., Schlesinger, S., and Boeing, H. (2017). Food groups and risk of hypertension : a systematic review and dose-response meta-analysis of prospective studies. *Advances in Nutrition*, 8(6) :793–803.

[40] Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso : Variable selection for causal inference. *Biometrics*, 73(4) :1111–1122.

[41] Talbot, D., Lefebvre, G., and Atherton, J. (2015). The Bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2) :207–236.

[42] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398) :528–540.

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

[44] Van der Laan, M. J. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

[45] Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1) :7–30.

[46] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE transactions on information theory*, 55(5) :2183–2202.

[47] Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479) :1039–1048.

[48] Wang, X. (2006). Approximating bayesian inference by weighted likelihood. *The Canadian Journal of Statistics/La revue canadienne de statistique*, pages 279–298.

[49] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov) :2541–2563.

[50] Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1) :263–273.

[51] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.