UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE MULTIDIMENSIONNELLE DE L'IMPACT DES TRAITEMENTS SUR LA PRÉVALENCE ET LA SÉVÉRITÉ DES SÉQUELLES À LONG TERME CHEZ LES SURVIVANTS DE LEUCÉMIE AIGUË LYMPHOBLASTIQUE DE L'ENFANT

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MIGUEL CAUBET FERNÁNDEZ

DÉCEMBRE 2017

REMERCIEMENTS

Je voudrais tout d'abord remercier ma directrice de recherche Geneviève Lefebvre qui m'a fait confiance et qui m'a encouragé à réaliser ce projet de recherche. Merci Dr Lefebvre pour votre disponibilité, votre soutien et votre aide précieuse pendant le déroulement de ce mémoire. Votre sens de l'engagement et votre esprit critique et constructif resteront comme un exemple à suivre pour moi. J'aimerais aussi remercier Mariia Samoilenko pour ses remarques toujours pertinentes et pour avoir partagé ses compétences avec autant de passion.

Je tiens à adresser mes remerciements les plus sincères à Dr Daniel Sinnett pour la disponibilité des données et son appui financier. Je tiens également à remercier Dr Simon Drouin pour son accueil chaleureux et son encadrement lors de mon stage au Centre de recherche du CHU Sainte-Justine. Merci aussi aux différents membres du laboratoire qui, à un moment ou à un autre, m'ont aidé à mieux comprendre le projet PÉTALE et la nature des données collectées.

Je voudrais remercier les professeurs du Département de mathématiques de l'UQÀM pour leur contribution à ma formation, particulièrement Professeure Sorana Froda pour son encadrement au début de mes études de maîtrise. Merci également à Gisèle Legault pour sa disponibilité et son support au Laboratoire d'informatique des cycles supérieurs en mathématiques.

Finalement, je dédie ce mémoire à mes parents et à ma fille Marina. C'est grâce à votre amour inconditionnel que j'ai eu la force de me relever et de continuer pendant ces deux dernières années.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Dans le cadre de ce mémoire, on s'intéresse à mieux comprendre le lien entre le traitement de la leucémie aiguë lymphoblastique de l'enfant (LALe) et le développement de certaines complications médicales à long terme. En particulier, on vise à estimer l'effet combiné de l'exposition à la radiothérapie (RT) et des doses reçues de corticostéroïdes (CS) sur le risque de présenter quatre troubles cardiométaboliques : obésité, résistance à l'insuline, (pré)-hypertension et dyslipidémie. À cet effet, on dispose d'un ensemble de données caractérisant une cohorte de 180 jeunes survivants de la LALe traités au Centre hospitalier St-Justine (CHUSJ) à Montréal. Compte tenu des possibles liens de dépendance entre les troubles cardiométaboliques, on implémente le modèle de régression logistique multivarié bayésien proposé par O'Brien et Dunson. En catégorisant RT et CS en deux groupes, on définit une variable d'exposition à trois niveaux. D'ailleurs, afin de réduire le biais dans l'estimation des effets, un ensemble de covariables potentiellement confondantes est inclus dans le modèle de régression. Des analyses de sensibilité sont également effectuées. Les résultats montrent que le niveau de traitement le plus élevé (versus le traitement de base) augmente le risque de présenter une dyslipidémie ($OR = 2.49$ (IC à 95% : $1.16, 5.48$)) et que ce même niveau de traitement est associé à une augmentation du risque de 0.15 (IC à 95% : $0.01, 0.30$) de subir au moins une complication cardiométabolique. On constate aussi que les mesures d'effet associées au traitement pour l'obésité et la résistance à l'insuline sont sensibles à l'ajustement pour la concentration de cellules blanches au diagnostic (WBC). On conclut que le niveau de traitement le plus élevé est associé significativement à la dyslipidémie et qu'il ne semble pas avoir un impact sur le risque de présenter plusieurs troubles cardiométaboliques en même temps. Enfin, la sévérité de la maladie au diagnostic, mesurée par WBC, pourrait être la cause du risque accru d'être insulinorésistant chez les jeunes survivants de la LALe.

MOTS-CLÉS : Leucémie aiguë lymphoblastique de l'enfant, troubles cardiométaboliques, effets du traitement du cancer, données binaires corrélées, statistique bayésienne, algorithme MCMC.

INTRODUCTION

La leucémie aiguë lymphoblastique de l'enfant (LALe) est le cancer pédiatrique le plus commun, étant la cause d'environ 25% des cas de cancer chez ceux-ci. Au cours des dernières années, des progrès dans la compréhension de la pathobiologie de la LALe ont mené à l'élaboration de traitements adaptés au risque des patients permettant d'améliorer considérablement les chances de guérison à long terme ($> 85\%$) (Pui *et al.*, 2012). Cependant, pour un grand nombre de survivants, l'exposition à la radiothérapie (RT) et plusieurs agents chimiothérapeutiques pendant une période vulnérable de leur vie a des répercussions importantes à long terme. Aujourd'hui, il est connu que les survivants de la LALe constituent un groupe à haut risque de développer différentes complications médicales associées au traitement, avec une incidence cumulée de maladies chroniques dépassant le 60% (Haddy *et al.*, 2009; Mody *et al.*, 2008).

Parmi les complications médicales les plus communes chez les survivants de la LALe, on trouve des troubles cardiométaboliques tels que l'obésité, la résistance à l'insuline, la (pré)-hypertension et la dyslipidémie. Puisque l'apparition de complications cardiométaboliques implique une perte importante de la qualité de vie des survivants, ainsi qu'un risque accru de développer des complications cardiovasculaires graves, il est pertinent de comprendre l'impact du traitement sur le risque de les développer. Dans des travaux antérieurs, la RT cranienne a été identifiée comme le plus grand facteur de risque pour l'obésité et a aussi été associée à une diminution de la sensibilité à l'insuline ainsi qu'à une augmentation de la masse grasse (Bulow *et al.*, 2004). Par ailleurs, les corticostéroïdes (CS), la classe des agents chimiothérapeutiques qui constitue l'épine dorsale du traitement, ont

été associés à une augmentation du poids des patients (Lughetti *et al.*, 2012), ainsi qu'à des troubles neurocognitifs et des complications osseuses (Waber *et al.*, 2000; Padhye *et al.*, 2016). Il se peut donc que la combinaison de CS et RT dans le traitement puisse contribuer à la détérioration de la santé cardiométabolique des patients selon plusieurs mécanismes, comme par exemple, en endommageant les organes endocriniens et en dégradant la fonction endothéliale et le métabolisme du tissu adipeux. Néanmoins, bien qu'un grand nombre d'études se sont penchées sur le lien entre le traitement de la LALe et l'apparition des effets indésirables à long terme (EILs), celui-ci n'a pas été bien analysé dans des groupes de jeunes survivants (enfants, adolescents et jeunes adultes). De plus, dans les dernières décennies, des changements dans les protocoles de traitement de la LALe ont mené à une réduction de l'exposition à la RT et à une diminution des doses de radiation, soulevant ainsi la question à savoir si les patients traités avec les thérapies plus actuelles sont encore à risque de subir des complications cardiométaboliques.

Motivé par ces questions, ce mémoire s'intéresse à estimer les effets combinés de RT et doses de CS sur le risque de développer quatre complications cardiométaboliques, obésité, résistance à l'insuline, (pré)-hypertension et dyslipidémie, dans une cohorte de jeunes survivants de la LALe traités au Centre hospitalier St-Justine (CHUSJ) à Montréal (projet PÉTALE). Étant donné les potentielles relations de dépendance entre ces quatre réponses cardiométaboliques, on a décidé de mettre en œuvre un modèle statistique multivarié pouvant tenir compte de cette structure de dépendance. À cet effet, le modèle de régression logistique multivarié proposé par O'Brien et Dunson (2004) apparaît particulièrement indiqué à cause de la grande interprétabilité des résultats générés : les coefficients de régression du modèle peuvent être interprétés comme des logarithmes de rapports de cotes (de la même façon que dans la régression logistique univariée standard), en même temps que la structure de dépendance entre les réponses est modélisée

de façon simple et sans restriction. En outre, ce modèle est spécialement adapté pour l'inférence bayésienne, permettant d'estimer efficacement l'impact du traitement sur l'ensemble des complications cardiométaboliques, marginalement et conjointement, avec plusieurs mesures d'association d'intérêt.

Ainsi, dans cette étude, on a implanté le modèle de O'Brien et Dunson (2004) pour identifier les niveaux de traitement associés aux EILs cardiométaboliques à partir d'une cohorte de 180 survivants provenant de l'étude PÉTALE menée au CHUSJ entre 2012 et 2016. La compréhension des liens entre le traitement et ces complications médicales pourront éventuellement contribuer à l'identification de biomarqueurs précoces d'apparition des EILs, permettant ainsi d'élaborer des lignes directrices pour le suivi personnalisé des survivants de la LALe.

CHAPITRE I

LA RÉGRESSION LOGISTIQUE MULTIVARIÉE BAYESIENNE

1.1 Motivation

Un choix tout à fait naturel pour l'analyse associationnelle entre le traitement de la LAL chez l'enfant et l'apparition de troubles cardiométaboliques est le modèle de régression logistique multiple standard. Dans ce contexte, les analyses sont effectuées pour chaque trouble métabolique de façon indépendante aux autres, c'est-à-dire en considérant chaque trouble binaire comme une variable réponse dans une suite de modèles de régression logistique séparés. Or, il est raisonnable de penser que, pour un même individu, l'apparition d'un trouble métabolique n'est pas indépendante de l'apparition des autres troubles. Ainsi, il nous semble pertinent d'étudier la relation entre les traitements et les différentes réponses métaboliques à l'aide d'un modèle de régression multivariée.

À cet effet, on a considéré avantageux de suivre une approche bayésienne pour l'analyse de nos données, au détriment d'un approche fréquentiste. En effet, les raisons pour l'utilisation d'une approche bayésienne sont multiples, notamment la petite taille de notre échantillon, la possibilité d'introduire de l'information *a priori* sur les paramètres du modèle et la facilité avec laquelle l'inférence peut être effectuée pour diverses mesures de risque.

Parmi les méthodes de régression les plus populaires pour l'analyse multivariée

de données binaires, on trouve les équations d'estimation généralisées (GEEs ; Zeger et Liang,1986) et les modèles linéaires généralisés à effets mixtes (Stiralli *et al.*, 1984). Bien que le modèle GEE a l'avantage de préserver l'interprétation marginale ("population-averaged") des coefficients de la régression ainsi que d'être robuste aux erreurs de spécification dans la structure de corrélation, il n'est pas adapté pour l'inférence bayésienne car ne nécessitant pas la spécification d'une vraisemblance. En ce qui concerne les modèles à effets mixtes, on identifie plusieurs problèmes. Premièrement, lorsqu'on utilise un modèle à effets mixtes bayésien, les coefficients de la régression s'interprètent comme des effets sur l'individu moyen ; ils perdent donc leur interprétation marginale, qui est plus intuitive et facile à vulgariser. Deuxièmement, il a été montré que des distributions simples non-informatives pour la covariance (ou paramètres de la covariance) génèrent des distributions *a posteriori* impropres (Natarajan et Kass, 2000). Finalement, la présence d'effets mixtes complexifie les calculs pour l'obtention de la loi *a posteriori* des paramètres d'intérêts.

Une alternative intéressante aux méthodes mentionnées précédemment consiste en l'utilisation de modèles probit bayésiens. Ces modèles peuvent être spécifiés à partir de variables auxiliaires, ce qui permet d'améliorer l'efficacité computationnelle de l'échantillonnage des lois *a posteriori* des paramètres du modèle (Albert et Chib, 1993). Un autre avantage du modèle probit est que la structure de dépendance entre les variables binaires est spécifiée de façon simple et flexible à partir de la matrice de corrélation des variables auxiliaires normales. Par contre, l'interprétation des coefficients dans la régression est moins intuitive que dans les modèles logistiques.

O'Brien et Dunson (2004), inspirés par les avantages des modèles probit par rapport aux modèles à effets mixtes, ont proposé un modèle de régression bayésien spécifié à partir de variables auxiliaires logistiques. Ce modèle conserve tous les

avantages des modèles probit bayésiens, en plus de préserver la structure marginale logistique des coefficients de régression, c'est-à-dire, en permettant l'interprétation directe des résultats sous forme de rapports de cotes. Ainsi, le modèle logistique bayésien de O'Brien et Dunson (2004) est considéré comme un modèle de choix pour l'analyse de nos données. Dans les prochaines sections, on introduit les bases théoriques de ce modèle.

1.2 Modèle logistique multivarié de O'Brien et Dunson (2004)

Le modèle de régression logistique multivarié proposé par O'Brien et Dunson (2004) spécifie la vraisemblance des variables binaires à l'aide de variables auxiliaires (ou latentes) distribuées selon une loi logistique multivariée. Cette loi, à la différence d'autres proposées antérieurement (Gumbel, 1961; Malik et Abraham, 1973; Castillo *et al.*, 1997), possède une structure de corrélation sans contraintes. Ainsi, la structure de dépendance entre les différentes variables binaires peut être modélisée de façon flexible à partir des coefficients de corrélation de la loi logistique proposée.

Dans ce qui suit, on présente la façon dont les variables auxiliaires aléatoires logistiques peuvent être utilisées pour spécifier des modèles de régression logistique univarié et multivarié.

1.2.1 Variables latentes et régression logistique

Supposons, d'abord, le modèle de régression logistique univarié suivant :

$$\text{logit}\, \mathrm{P}(Y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i' \boldsymbol{\beta}, \tag{1.1}$$

où Y_i est une variable réponse binaire, $\boldsymbol{x}_i$ est un vecteur de variables explicatives et $\boldsymbol{\beta}$ est le vecteur des coefficients inconnus de la régression. Si, plutôt, on exprime

$\mathrm{P}(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta})$ en fonction de $\boldsymbol{x}_i'\boldsymbol{\beta}$, on retrouve l'autre façon commune de présenter le lien entre la variable réponse et les variables explicatives dans ce type de régression :

$$\mathrm{P}(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{H}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})},$$

où $\mathrm{H}(x) = \frac{\exp(x)}{1+\exp(x)}$ est la fonction de répartition d'une variable logistique standard (moyenne 0 et paramètre d'échelle 1). De façon générale, la densité d'une variable logistique de moyenne μ et paramètre d'échelle s est de la forme :

$$\mathcal{L}(z|\mu, s) = \frac{\exp(-\frac{z-\mu}{s})}{s(1 + \exp(-\frac{z-\mu}{s}))^2}, \text{ avec } z \in (-\infty, \infty).$$

Ainsi, un modèle équivalent au modèle (1.1) est :

$$\begin{cases} Y_i = \mathbb{1}(Z_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}) \\ Z_i \sim \mathcal{L}(0, 1), \end{cases} \tag{1.2}$$

où $\mathbb{1}(\cdot)$ est la fonction indicatrice et Z_i est une variable auxiliaire suivant une loi logistique standard.

Remarque. Dans ce mémoire, on utilise la notation $\mathcal{L}(\mu, s)$ pour désigner une variable aléatoire logistique de moyenne μ et de paramètre d'échelle s et la notation $\mathcal{L}(\cdot|\mu, s)$ pour désigner la densité correspondante.

À la base du modèle de O'Brien et Dunson (2004) est cette formulation équivalente à celle de (1.2) où la relation entre la variable auxiliaire logistique et la réponse ne dépend pas des valeurs de $\boldsymbol{\beta}$

$$\begin{cases} Y_i = \mathbb{1}(Z_i > 0) \\ Z_i \sim \mathcal{L}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1). \end{cases} \tag{1.3}$$

Pour vérifier l'équivalence entre les deux formulations, on définit $Z_i \sim \mathcal{L}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ et $Z_i^0 \sim \mathcal{L}(0,1)$. Alors, par les propriétés d'invariance par translation et de symétrie de la loi logistique, on a

$$\begin{aligned} \mathrm{P}(Z_i > 0) &= \mathrm{P}(Z_i - \boldsymbol{x}_i'\boldsymbol{\beta} > -\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \mathrm{P}(Z_i^0 > -\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \mathrm{P}(Z_i^0 \leq \boldsymbol{x}_i'\boldsymbol{\beta}). \end{aligned}$$

Lorsqu'on se situe dans le cas multivarié, la réponse associée à la $i^{\text{ème}}$ unité consiste en un vecteur de variables binaires $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ potentiellement corrélées. Une façon de spécifier un modèle de régression logistique qui peut tenir compte de la structure de dépendance entre ces variables consiste à généraliser la spécification du modèle univarié avec variables auxiliaires. Ainsi, soit $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ un vecteur aléatoire distribué selon une loi multivariée dont les lois marginales sont logistiques standards de moyenne $\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j$ $(j = 1, \ldots, p)$. En posant $Y_{ij} = \mathbb{1}(Z_{ij} > 0)$, on relie chaque variable Y_{ij} au prédicteur linéaire $\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j$ selon une régression logistique univariée, en même temps qu'on modélise la dépendance entre les variables réponses au moyen de la structure de corrélation de la loi logistique multivariée. Or, tel que mentionné précédemment, plusieurs lois multivariées logistiques rencontrées dans la littérature manquent de flexibilité du point du vue de la structure de corrélation, ce qui restreint leur utilisation dans l'approche de modélisation multivariée décrite plus haut. Dans le but de résoudre cette problématique, O'Brien et Dunson (2004) ont proposé une nouvelle loi multivariée logistique plus flexible.

Dans la prochaine section, on présente quelques résultats reliés à la loi de Student multivariée qui permettent de mieux comprendre le modèle proposé qui est décrit dans la section 1.2.3.

1.2.2 La loi de Student multivariée

Une loi de Student multivariée est une généralisation aux vecteurs aléatoires de la loi de Student univariée. Bien qu'il existe différentes possibilités pour généraliser la loi de Student univariée, on étudie dans cette section la généralisation la plus répandue dans la littérature. Cette loi multivariée est celle utilisée pour bâtir le modèle de régression logistique de O'Brien et Dunson (2004).

Définition 1.2.1. Soit $V > 0$ une variable aléatoire qui suit une loi gamma $\text{Gam}(\nu/2, \nu/2)$ avec $\nu > 0$ et $\boldsymbol{W}$ un vecteur aléatoire de dimension p indépendant de V normalement distribué $\mathcal{N}(0, \boldsymbol{\Sigma})$, (où $\boldsymbol{\Sigma}$ est la matrice de covariances de $\boldsymbol{W}$). Le vecteur aléatoire $\boldsymbol{T}$ défini comme :

$$\boldsymbol{T} = \boldsymbol{\mu} + \frac{1}{\sqrt{V}}\boldsymbol{W}, \tag{1.4}$$

où $\boldsymbol{\mu}$ est un vecteur $p \times 1$, suit une loi de Student multivariée.

On note que la matrice de covariances de la loi de Student multivariée est définie seulement lorsque $\nu > 2$ et dans ce cas, on a $\text{Cov}(\boldsymbol{T}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$.

Proposition 1.2.1. Sous l'hypothèse que la matrice de covariances $\boldsymbol{\Sigma}$ du vecteur $\boldsymbol{W}$ soit inversible, le vecteur $\boldsymbol{T}$ possède une fonction de densité

$$\mathcal{T}_p(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left(1 + \frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right)^{-(\nu+p)/2}.$$

Démonstration. Puisque V et $\boldsymbol{W}$ sont indépendantes, leur densité conjointe est le produit de leur densités individuelles

$$f_{V,\boldsymbol{W}}(v, \boldsymbol{w}) = f_V(v) f_{\boldsymbol{W}}(W).$$

En introduisant la paire de variables aléatoires

$$\boldsymbol{T} = \boldsymbol{\mu} + \frac{1}{\sqrt{V}}\boldsymbol{W} \qquad\qquad U = V,$$

où $\boldsymbol{T}$ est notre vecteur aléatoire d'intérêt et U est une variable aléatoire auxiliaire, on peut appliquer le théorème de changement de variables dans le but d'obtenir la densité conjointe de $\boldsymbol{T}$ et U. Plus spécifiquement, on pose :

$$\begin{aligned} V &= U \\ \boldsymbol{T} &= \boldsymbol{\mu} + \frac{1}{\sqrt{V}}\boldsymbol{W} \rightarrow \boldsymbol{W} = \sqrt{U}(\boldsymbol{T} - \mu). \end{aligned}$$

Puis, en effectuant le changement de variables, on obtient :

$$\begin{aligned} f_{U,\boldsymbol{T}}(u, \boldsymbol{t}) &= f_V(u) f_{\boldsymbol{W}}(\sqrt{u}(\boldsymbol{t} - \mu))||\boldsymbol{J}|| \\ &= \text{Gam}(u|\nu/2, \nu/2)\mathcal{N}(\sqrt{u}(\boldsymbol{t} - \mu)|\mu, \boldsymbol{\Sigma})u^{p/2} \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{(\nu+p)/2-1} \exp\left(-\frac{\nu}{2}u\right) \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{\Delta^2}{2}u\right), \end{aligned}$$

où $||\boldsymbol{J}|| = u^{p/2}$ est la valeur absolue du déterminant jacobien $|\boldsymbol{J}|$ de la transformation et $\Delta = \sqrt{(\boldsymbol{t} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{t} - \boldsymbol{\mu})}$ est la distance de Mahalanobis.

Finalement, la marginalisation de cette densité par rapport à la variable auxiliaire U, nous donne la densité du vecteur $\boldsymbol{T}$:

$$\begin{aligned} f_{\boldsymbol{T}}(\boldsymbol{t}) &= \int_0^\infty f_{U,\boldsymbol{T}}(u, \boldsymbol{t})\, \mathrm{d}u \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \int_0^\infty u^{\frac{\nu+p}{2}-1} \exp\left(-\frac{\nu + \Delta^2}{2}u\right) \mathrm{d}u \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left(\frac{\nu + \Delta^2}{2}\right)^{-(\nu+p)/2} \Gamma\left(\frac{\nu + p}{2}\right) \\ &= \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left(1 + \frac{1}{\nu}(\boldsymbol{t} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{t} - \boldsymbol{\mu})\right)^{-(\nu+p)/2}. \end{aligned}$$

□

À partir de maintenant, on note cette densité par $\mathcal{T}_p(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ où ν correspond au nombre de degrés de liberté. On rappelle que, pour garantir l'existence de la

densité $\mathcal{T}_p(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ du vecteur $\boldsymbol{T}$, la matrice $\boldsymbol{\Sigma}$ (appelée matrice d'échelle) doit être définie positive.

Remarque. Bien que dans notre définition du vecteur aléatoire $\boldsymbol{T}$ on a utilisé une variable $V \sim \text{Gam}(\nu/2, \nu/2)$, il est possible de montrer qu'une densité de la même famille peut être obtenue dans le cas plus général où $V \sim \text{Gam}(\alpha/2, \beta/2)$, $\alpha > 0, \beta > 0$.

Une propriété intéressante de loi Gamma est que lorsque $\alpha = \beta \to \infty$, sa densité s'approche de l'impulsion de Dirac $\delta(u - 1)$. Ainsi, en combinant ce résultat avec l'expression (1.4), on déduit que lorsque le nombre de degrés de liberté ν tend vers l'infini, $V = 1$ avec probabilité 1 et $\boldsymbol{T}$ est normalement distribuée. Par conséquent, cette loi de Student peut être vue comme une généralisation de la loi normale multivariée et certaines de ses propriétés sont communes à la loi normale.

Dans le contexte de notre recherche, on considère pertinent de mentionner la propriété suivante. Soit $\boldsymbol{T}$ un vecteur aléatoire de densité $\mathcal{T}_p(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, $\boldsymbol{A}$ une matrice $p_z \times p$ et $\boldsymbol{b}$ un vecteur de dimension p_z. Alors, sous la condition que la matrice $\boldsymbol{A\Sigma A^T}$ soit inversible, le vecteur $\boldsymbol{Z} = \boldsymbol{AT} + \boldsymbol{b}$ a comme densité

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \mathcal{T}_{p_z}(\boldsymbol{z}|\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A^T}, \nu).$$

Ce résultat est facilement vérifié à partir de l'expression (1.4) et de la formule de transformation linéaire d'une variable normale multivariée. Une conséquence directe de ce résultat est que si le vecteur $\boldsymbol{T}$ suit une loi de Student multivariée, alors tout sous-vecteur de $\boldsymbol{T}$ est aussi de loi de Student. Pour le voir, il suffit de prendre $\boldsymbol{T^T} = (\boldsymbol{T_1^T T_2^T})$ et d'appliquer le résultat précédent avec la matrice $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}]$.

Une caractéristique bien connue de la loi normale multivariée est que la matrice de corrélation détermine totalement la dépendance ou, autrement dit, si deux

vecteurs normaux ne sont pas corrélés, alors ils sont indépendants. Par contre, ceci n'est plus vrai dans le cas de la loi de Student multivariée. Pour voir ce résultat, on considère un vecteur $\boldsymbol{T}^T = [\boldsymbol{T}_1^T\, \boldsymbol{T}_2^T] \sim \mathcal{T}_p(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ tel que $\boldsymbol{T}_1$ et $\boldsymbol{T}_2$ ne sont pas corrélés. Alors, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_{22}} \end{pmatrix}$ et la densité prend la forme

$$f_{\boldsymbol{T}_1,\boldsymbol{T}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) \propto \left(1 + \frac{1}{\nu} \begin{pmatrix} \boldsymbol{t}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{t}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_{22}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{t}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{t}_2 - \boldsymbol{\mu}_2 \end{pmatrix}\right)^{-(\nu+p)/2}$$

$$\propto \left(1 + \frac{1}{\nu}(\boldsymbol{t}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma_{11}}^{-1}(\boldsymbol{t}_1 - \boldsymbol{\mu}_1) + \frac{1}{\nu}(\boldsymbol{t}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma_{22}}^{-1}(\boldsymbol{t}_2 - \boldsymbol{\mu}_2)\right)^{-(\nu+p)/2}.$$

Puisqu'il n'est pas possible de factoriser cette densité en deux termes, l'un dépendant de $\boldsymbol{t}_1$ et l'autre de $\boldsymbol{t}_2$, on déduit que les vecteurs $\boldsymbol{T}_1$ et $\boldsymbol{T}_2$ ne sont pas indépendants.

1.2.3 La distribution logistique multivariée du modèle de O'Brien et Dunson (2004)

Une façon d'obtenir une loi logistique multivariée consiste à appliquer une transformation simple aux vecteurs qui suivent une loi multivariée standard, comme par exemple la loi normale ou la loi de Student. Pour le voir, supposons que le vecteur $\boldsymbol{E} = (E_1, \ldots, E_p)$ suit une loi multivariée continue sans paramètre de position ni d'échelle, et soit F la fonction de répartition marginale (univariée) des E_j. D'abord, on utilise le résultat bien connu que $F(E_j) \sim \mathcal{U}(0, 1)$.

Démonstration. Par hypothèse la loi est continue, ce qui implique F est une fonction de répartition continue. En supposant de plus que F est strictement croissante, ce qui est le cas pour les lois les plus communes, elle admet une inverse F^{-1}

qui est strictement croissante. Ainsi, si $0 \leq v \leq 1$

$$\begin{aligned} \mathrm{P}(F(E_j) \leq v) &= \mathrm{P}(F^{-1}(F(E_j)) \leq F^{-1}(v)) \\ &= \mathrm{P}(E_j \leq F^{-1}(v)) \\ &= F(F^{-1}(v)) = v. \end{aligned}$$

□

Posons $Z = \log\left(\frac{U}{1-U}\right)$, où $\mathrm{U} \sim \mathcal{U}(0,1)$. En appliquant le changement de variable $u = \frac{\exp(z)}{1+\exp(z)}$ sur la densité $\mathrm{f}_U(u) = \mathbb{1}(0 \leq u \leq 1)$ on obtient $\mathrm{f}_Z(z) = \frac{\exp(z)}{(1+\exp(z))^2}$, ce qui correspond à la densité d'une variable logistique standard. Ainsi, les variables définies comme $Z_j = \mu_j + \log\left(\frac{F(E_j)}{1-F(E_j)}\right)$, $j = 1, \ldots, p$, sont des variables logistiques de moyennes μ_j, et le vecteur aléatoire $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ est distribué selon une loi logistique multivariée de moyenne $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$.

Dans un premier temps, on pourrait penser qu'une option simple et raisonnable comme choix de loi multivariée pour $\boldsymbol{E}$ est la loi normale multivariée. Par contre, on verra subséquemment que le choix de la loi de Student multivariée est intéressant puisque celui-ci nous permet d'approcher ce modèle logistique par un autre plus réalisable du point du vue computationnel.

Par ailleurs, mentionnons que l'utilisation de la loi de Student dans la définition de la loi logistique multivariée permet de paramétrer cette dernière en fonction du vecteur de moyennes $\boldsymbol{\mu}$, une matrice de corrélation $\boldsymbol{R}$ et un nombre de degrés de liberté ν. On se souvient que la densité de la loi de Student a été spécifiée à partir de la matrice de covariances $\boldsymbol{\Sigma}$; or, pour que le modèle soit identifiable, on doit restreindre cette matrice à l'espace des matrices de corrélation (Chib et Greenberg, 1998).

Dans notre analyse de régression, chaque composante du vecteur $\boldsymbol{\mu}$ est un prédicteur linéaire associé à une réponse binaire Y_j, et les éléments de $\boldsymbol{R}$, les corrélations

entre les composantes du vecteur Student, nous donnent une mesure de la dépendance entre les Y_j. En ce qui concerne le paramètre ν, on décrit son rôle dans l'approximation de notre modèle dans la section 1.4.

Pour la suite de la présentation, on dérive l'expression de la densité logistique multivariée induite par l'utilisation d'une Student multivariée.

Soit $\boldsymbol{T} = (T_1, \ldots, T_p)$ un vecteur distribué selon une loi de Student multivariée avec ν degrés de liberté, moyenne $\mathbf{0}$ et matrice d'échelle $\boldsymbol{R}$. La densité d'un tel vecteur est

$$\mathcal{T}_p(\boldsymbol{t}; \mathbf{0}, \boldsymbol{R}, \nu) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\boldsymbol{R}|^{1/2}} \left(1 + \frac{1}{\nu}\boldsymbol{t}'\boldsymbol{R}^{-1}\boldsymbol{t}\right)^{-(\nu+p)/2}.$$

Dans la section 1.2.2, on a vu que tout sous-vecteur d'un vecteur Student est aussi distribué selon une loi de Student. Ainsi, on déduit que les lois marginales univariées associées aux composantes T_j, $j = 1, \ldots, p$, de $\boldsymbol{T}$ sont des variables Student univariées standard de moyenne 0 et ν degrés de liberté. Alors, en notant par F_ν la fonction de répartition d'une telle loi, le vecteur $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ est défini à partir de la transformation

$$Z_j = \mu_j + \log\left(\frac{F_\nu(T_j)}{1 - F_\nu(T_j)}\right), \quad j = 1, \ldots, p.$$

Par la suite, on applique le changement de variables $t_j = F_\nu^{-1}\left(\dfrac{\exp(z_j - \mu_j)}{1 + \exp(z_j - \mu_j)}\right)$.

En remarquant que $(F_\nu^{-1})'(x) = \dfrac{1}{F_\nu'(F_\nu^{-1}(x))} = \dfrac{1}{\mathcal{T}(F_\nu^{-1}(x)|0,1,\nu)}$,

on obtient

$$\frac{\partial t_j}{\partial z_k} = \begin{cases} \dfrac{\mathcal{L}(z_j|\mu_j, 1)}{\mathcal{T}(g_\nu(z_j - \mu_j)|0,1,\nu)} & \text{si } k = j \\ 0 & \text{sinon} \end{cases},$$

où $g_\nu(x) = F_\nu^{-1}(\exp(x)/(1+\exp(x)))$.

Ainsi, on a $||\boldsymbol{J}|| = \prod_{j=1}^{p} \frac{\mathcal{L}(z_j|\mu_j,1)}{\mathcal{T}(g_\nu(z_j-\mu_j)|0,1,\nu)}$ et la densité du vecteur $\boldsymbol{Z}$ prend la forme suivante

$$\begin{aligned}\mathcal{L}_p(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{R},\nu) = & \mathcal{T}_p((g_\nu(z_1-\mu_1),\ldots,g_\nu(z_p-\mu_p))'|\boldsymbol{0},\boldsymbol{R}) \\ & \times \prod_{i=1}^{p} \frac{\mathcal{L}(z_j|\mu_j,1)}{\mathcal{T}(g_\nu(z_j-\mu_j)|0,1,\nu)}.\end{aligned}$$

Dans un contexte de régression, pour chaque individu ($i = 1,\ldots,n$) on a p réponses binaires Y_{ij} et, associé à chaque réponse, un prédicteur linéaire $\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j$. Ainsi, la moyenne $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)$ de la distribution logistique multivariée correspond au produit $\boldsymbol{X}_i\boldsymbol{\beta}$, où $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1,\ldots,\boldsymbol{\beta}'_p)$, $\boldsymbol{\beta}_j \in \mathbb{R}^{q_j}$ est un vecteur de paramètres, et $\boldsymbol{X}_i = \mathrm{diag}(\boldsymbol{x}'_{i1},\ldots,\boldsymbol{x}'_{ip})$ est une matrice de covariables de dimension $p \times q \quad (q = \sum q_j)$.

De cette façon, associée à la $i^{\text{ème}}$ observation on a la relation suivante entre les vecteurs aléatoires $\boldsymbol{Z}$ et $\boldsymbol{Y}$:

$$\begin{aligned} Y_{ij} &= \mathbb{1}(Z_{ij} > 0) \quad j = 1,\ldots,p, \\ \boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{ip}) &\sim \mathcal{L}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{R},\nu). \end{aligned}$$

La vraisemblance de $\boldsymbol{Y}_i$ est donc

$$\begin{aligned}&\mathrm{P}(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\beta},\boldsymbol{R}) = \\ &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\prod_{j=1}^{p}\mathbb{1}(z_{ij}>0)^{y_{ij}}\mathbb{1}(z_{ij}\le 0)^{1-y_{ij}}\right)\mathcal{L}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{R},\nu)\,\mathrm{d}\boldsymbol{z}_i. \quad (1.5)\end{aligned}$$

1.3 La statistique bayésienne et les méthodes de simulation Monte-Carlo par chaînes de Markov (MCMC)

Dans cette section, on introduit les fondements de l'inférence bayésienne, ainsi que la technique de simulation utilisée pour obtenir des échantillons de la loi *a posteriori* des paramètres du modèle multivarié étudié.

1.3.1 L'approche bayésienne pour l'estimation de paramètres

Dans un contexte d'inférence, on s'intéresse à tirer des conclusions sur deux classes de quantités non observées : les quantités qui peuvent être potentiellement observées (des observations futures d'un processus, des résultats contrefactuels en inférence causale, etc.) et des quantités qui ne sont pas directement observables comme c'est le cas pour les paramètres d'un modèle.

Le principe fondamental de l'inférence bayésienne consiste à quantifier, en termes probabilistes, notre état de connaissances sur des quantités non observées associées à un hypothétique processus générateur de données (modèle). Ainsi, à la différence de l'approche fréquentiste, où les paramètres à estimer sont traités comme des constantes inconnues, l'approche bayésienne traite ces derniers comme des variables aléatoires. Plus précisément, on utilise des lois de probabilités conditionnelles aux valeurs observées des données pour décrire l'état de nos connaissances sur ces quantités d'intérêt inconnues.

L'analyse statistique bayésienne des données peut être divisée en deux étapes principales : spécification du modèle et le calcul et interprétation de la loi *a posteriori.*

Spécification d'un modèle bayésien

Dans ce qui suit, on note par $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_d)$ le vecteur aléatoire des paramètres à estimer et par $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ le vecteur aléatoire associé aux données ob-

servables. Alors, le vecteur $\boldsymbol{Y}$ est formé par la concaténation des vecteurs $\boldsymbol{Y}_i$, $i = 1, \ldots, n$, où le vecteur $\boldsymbol{Y}_i$ est le vecteur aléatoire associé aux données issues de la $i^{\text{ème}}$ unité expérimentale.

Dans le but de pouvoir inférer sur les paramètres Θ à partir d'un échantillon de données $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$, on a besoin de spécifier un modèle consistant en une loi de probabilité conjointe pour Θ et $\boldsymbol{Y}$. Cette loi est exprimée comme le produit de deux éléments : la loi *a priori* $p(\boldsymbol{\theta})$ des paramètres et la loi conditionnelle associée à l'échantillon $p(\boldsymbol{y}|\boldsymbol{\theta})$. La loi *a priori* permet de spécifier l'état de nos connaissances sur les valeurs plausibles des paramètres avant l'expérience, alors que la loi conditionnelle modélise le comportement aléatoire des observations en fonction des valeurs fixées des paramètres. Lorsqu'on considère la fonction $p(\boldsymbol{y}|\boldsymbol{\theta})$ comme fonction de $\boldsymbol{\theta}$ pour des valeurs fixes de $\boldsymbol{y}$, on parle de la *fonction de vraisemblance* $l(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$. On remarque aussi que la loi $p(\boldsymbol{y}|\boldsymbol{\theta})$ résume toute l'information que les données fournissent sur le paramètre Θ.

Calcul et interprétation des lois *a posteriori*

À partir d'un modèle bayésien et la règle de Bayes, on dérive aisément la loi des paramètres conditionnelle aux valeurs observées des données, $p(\boldsymbol{\theta}|\boldsymbol{y})$

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{1.6}$$

où $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ ou $p(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$ selon si le vecteur Θ prend des valeurs continues ou discrètes.

La loi $p(\boldsymbol{\theta}|\boldsymbol{y})$, appelée loi *a posteriori* des paramètres, condense quantitativement nos connaissances sur les paramètres du modèle à la lumière des données observées et peut être interprétée comme une actualisation des nos connaissances *a priori*. Une formulation équivalente à l'expression (1.6) consiste à omettre le facteur $p(\boldsymbol{y})$

qui, ne dépendant pas de $\boldsymbol{\theta}$, est traité comme une constante de normalisation :

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}).$$

On remarque que la simplicité conceptuelle du paradigme bayésien lui fournit assez de flexibilité et de généralité pour pouvoir l'appliquer à des modèles complexes, contenant un grand nombre de paramètres à estimer, et pour lesquels l'approche fréquentiste n'est pas adaptée. Par contre, et à l'exception de certains cas particuliers, la dérivation analytique des lois *a posteriori* pour des modèles complexes peut devenir assez ardue, même impossible. Dans ces situations, des techniques de simulation doivent être implémentées dans le but de caractériser les lois *a posteriori* à partir d'échantillons aléatoires.

La génération d'échantillons de taille suffisamment grande permet de tracer des figures fournissant une information presque complète sur les lois visées tels que les histogrammes, graphiques de contour, ou nuages de points. De plus, des résumés statistiques générés par des logiciels informatiques (par exemple, R) permettent d'estimer toute mesure de position, dispersion ou de forme caractérisant la loi *a posteriori* jusqu'à un degré de précision voulu. D'autre part, le calcul d'intervalles de crédibilité à $100(\alpha/2)\%$ (ou régions de crédibilité), obtenus à partir des estimations des centiles, donne une mesure de l'incertitude *a posteriori* sur les valeurs plausibles d'un ou plusieurs paramètres du modèle.

Parmi les différentes techniques de simulation, l'une des plus populaires est la simulation par chaînes de Markov (aussi appelée simulation Monte-Carlo par chaînes de Markov, MCMC). Cette technique de simulation est celle qui a été utilisée pour l'obtention d'échantillons des lois *a posteriori* liées aux modèles appliqués dans le cadre de notre étude, et elle est l'objet d'une introduction théorique dans la section qui suit. La simulation par chaînes de Markov est une méthode générale de simulation qui est très souvent utilisée lorsque ce n'est pas possible (ou très in-

efficace au niveau du calcul) d'échantillonner d'une loi de probabilité directement. Dans ces situations, les méthodes MCMC permettent de générer des réalisations de façon itérative, de telle façon qu'à chaque étape du processus, la réalisation est distribuée selon une loi qui est de plus en plus proche de la loi visée, dans notre cas $p(\boldsymbol{\theta}|\boldsymbol{y})$. Plus précisément, la réalisation obtenue à l'itération t dépend des valeurs de celles générées antérieurement au temps $t-1, t-2, \ldots, 1$, à travers d'une *loi de transition* $T_t(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1})$ dont sa spécification est la clé pour garantir la convergence de la loi du processus au temps t ($t \to \infty$) vers $p(\boldsymbol{\theta}|\boldsymbol{y})$.

On souligne que les échantillons obtenus avec les méthodes MCMC ne sont pas indépendants, et que les premières itérations du processus peuvent être fortement influencées par la loi initiale utilisée. Par conséquent, lorsqu'on utilise des méthodes MCMC pour simuler de lois *a posteriori*, il est nécessaire de faire une évaluation critique des échantillons ainsi générés, en incluant une analyse de convergence, avant que quelqu'inférence statistique soit faite.

1.3.2 Méthodes de Monte-Carlo par chaînes de Markov

Dans cette section, on introduit les méthodes de Monte-Carlo par chaînes de Markov. En particulier, on présente l'algorithme Metropolis-Hastings qui est le fondement de la méthode de simulation implémentée dans notre étude.

Problématique

Dans l'approche bayésienne, la loi *a posteriori*

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \tag{1.7}$$

est l'élément essentiel pour faire des inférences sur les paramètres (ou fonctions des paramètres) du modèle. Ainsi, on peut être intéressé à estimer n'importe quelle caractéristique de la loi *a posteriori* comme par exemple : les moments, quan-

tiles, intervalles de crédibilité, etc. Toutes ces quantités peuvent être exprimées en termes d'espérances *a posteriori* de fonctions de $\boldsymbol{\theta}$. L'espérance *a posteriori* d'une fonction $f(\boldsymbol{\theta})$ est

$$\mathrm{E}[f(\boldsymbol{\theta})|\boldsymbol{y}] = \frac{\int f(\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}. \tag{1.8}$$

Or, dans de nombreux cas, le calcul analytique d'intégrales de type (1.8) s'avère très compliqué, même impossible. Une approche alternative pour l'évaluation de ces intégrales est l'intégration par la méthode de Monte-Carlo, qui est présentée à la section qui suit.

Intégration et méthode de Monte-Carlo

Soit $\boldsymbol{X}$ un vecteur aléatoire de dimension p suivant une loi de densité non-normalisée $\pi'(\cdot)$. On est intéressé à évaluer l'expression :

$$\mathrm{E}[f(\boldsymbol{X})] = \frac{\int f(\boldsymbol{x})\pi'(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int \pi'(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \tag{1.9}$$

pour une fonction d'intérêt $f(\cdot)$. On remarque que l'intégrale du dénominateur dans l'expression (1.9) n'est pas nécessairement égale à 1, puisqu'on suppose que la loi de $\boldsymbol{X}$ est connue à une constante près (i.e., $X \sim \pi = \pi'/k$) ; ceci est très souvent le cas dans l'analyse bayésienne où l'obtention de la constante de normalisation $k = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ peut être assez ardue. On tient aussi à remarquer que, par simplicité dans l'exposition, on a supposé que la variable $\boldsymbol{X}$ est un vecteur aléatoire continu avec densité $\pi(\boldsymbol{x})$. Or, les méthodes présentées ici s'appliquent à des cas plus généraux où $\boldsymbol{X}$ est un mélange de variables continues et discrètes. Bien qu'une notation propre à la théorie de la mesure permettrait de tenir en compte tous le cas possibles, on considère que ce n'est pas essentiel pour transmettre les idées principales de la méthode.

Schématiquement, l'intégration Monte-Carlo consiste à générer un échantillon de T observations $\{\boldsymbol{X}_t,\ t = 0, \ldots, T-1\}$ issues de la loi $\pi(\cdot)$, et à approcher la valeur de l'intégrale (1.9) par la moyenne empirique des réalisations

$$\mathrm{E}[f(\boldsymbol{X})] \approx \frac{1}{T} \sum_{t=0}^{T-1} f(\boldsymbol{X}_t).$$

Lorsque les réalisations sont indépendantes, et sous la condition que $\mathrm{E}[f(\boldsymbol{X})] < \infty$, on sait, par la loi des grands nombres, que la moyenne empirique converge avec probabilité 1 vers l'espérance. Ainsi, en augmentant la taille T de l'échantillon aléatoire on peut rendre l'approximation aussi précise que l'on veut.

Malheureusement, dans le contexte bayésien, les lois de probabilité à intégrer sont des lois *a posteriori* de dimension élevée pour lesquelles ce n'est souvent pas possible de générer directement des échantillons aléatoires indépendants. Cependant, les méthodes de Monte-Carlo par chaînes de Markov (MCMC) surmontent cette difficulté en permettant de simuler une suite de variables aléatoires (pas nécessairement indépendantes) qui converge en loi vers la loi visée (π).

Dans la section qui suit, on introduit de façon informelle les chaînes de Markov, pour ensuite présenter l'algorithme Metropolis-Hastings.

Les chaînes de Markov

Soit $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots\}$ une suite de vecteurs aléatoires telle que la loi de $\boldsymbol{X}_{t+1}$, conditionnellement aux valeurs précédentes du processus, dépend uniquement de $\boldsymbol{X}_t$

$$\mathrm{P}(\boldsymbol{X}_{t+1} | \boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t) = \mathrm{P}(\boldsymbol{X}_{t+1} | \boldsymbol{X}_t).$$

C'est-à-dire, pour une valeur fixée de $\boldsymbol{X}_t$, la valeur du prochain vecteur $\boldsymbol{X}_{t+1}$ ne dépend pas de la sous-suite $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t-1}\}$. Une suite qui possède cette

caractéristique est appelée chaîne de Markov. Dans ce contexte, l'ensemble des valeurs pouvant être prises par chaque vecteur est appelée l'espace d'états (S), et $\mathrm{P}(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$ est appelé le noyau de transition de la chaîne.

Il est possible de montrer que, sous certaines conditions de régularité, toute chaîne de Markov converge vers une loi limite $\pi(\cdot)$ indépendamment des conditions initiales (π_0) (Nummelin, 1984; Revuz, 1975). Ainsi, lorsque t augmente, les différentes valeurs prises par les vecteurs aléatoires $\boldsymbol{X}_t$ s'approchent de plus en plus de réalisations dépendantes issues la loi limite $\pi(\cdot)$. Plus précisément, si $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t, \ldots$ est une réalisation d'une chaîne de Markov respectant certaines conditions et $f(\boldsymbol{x})$ une fonction d'intérêt, on a

$$\boldsymbol{X}_t \xrightarrow[t\to\infty]{d} \boldsymbol{X} \sim \pi(\boldsymbol{x}); \qquad \frac{1}{t+1}\sum_{k=0}^{t} f(\boldsymbol{X}_t) \xrightarrow[t\to\infty]{p.s.} \mathrm{E}_\pi[f(\boldsymbol{X})].$$

Supposons, maintenant, qu'on est intéressé à générer des observations d'une loi $\pi(\cdot)$, mais qu'il n'existe pas de méthode directe pour le faire. Ainsi, si on est capable de construire une chaîne de Markov facile à simuler et dont la loi limite est $\pi(\cdot)$, on peut envisager d'estimer certaines caractéristiques de la loi visée $\pi(\cdot)$ pourvu que le nombre de valeurs simulées soit assez grand. L'algorithme Metropolis-Hastings, présenté dans la prochaine section, permet la construction de chaînes de Markov convergeant vers une loi limite $\pi(\cdot)$ spécifiée en avance.

L'algorithme Metropolis-Hastings

L'algorithme Metropolis-Hastings (algorithme M-H) (Metropolis *et al.*, 1953; Hastings, 1970) constitue un moyen simple, versatile et relativement efficace de construire une chaîne de Markov convergeant vers une loi limite $\pi(\cdot)$ spécifiée par l'utilisateur. À chaque temps t de cet algorithme, le prochain état du processus $\boldsymbol{X}_{t+1}$ est obtenu en générant une réalisation candidate $\boldsymbol{Y}$ issue d'une loi de proposition $q(\cdot|\boldsymbol{X}_t)$. On remarque que la loi de proposition peut être dépendante

de l'état actuel $\boldsymbol{X}_t$ de la chaîne. Ensuite, le point candidat $\boldsymbol{Y}$ est accepté avec probabilité $\alpha(\boldsymbol{X}_t, \boldsymbol{Y})$, où

$$\alpha(\boldsymbol{X}_t, \boldsymbol{Y}) = \min\left(1, \frac{\pi(\boldsymbol{Y})q(\boldsymbol{X}_t|\boldsymbol{Y})}{\pi(\boldsymbol{X}_t)q(\boldsymbol{Y}|\boldsymbol{X}_t)}\right). \tag{1.10}$$

Si le point candidat est accepté, alors le prochain état de la chaîne est $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$. Par contre, si le candidat n'est pas accepté, la chaîne ne change pas d'état, c'est-à-dire $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$. Ainsi, le fonctionnement de l'algorithme M-H peut être résumé avec le schéma algorithmique suivant :

- `Initialiser` $\boldsymbol{X}_0 \sim \pi_0$.
- `Pour` $t = 0, 1, 2, \ldots, T-1$:
  - `Générer` $\boldsymbol{Y}$ de la loi $q(\cdot|\boldsymbol{X}_t)$;
  - `Générer` U de la loi $\mathcal{U}(0,1)$;
  - `Si` $U \leq \alpha(\boldsymbol{X}_t, \boldsymbol{Y})$ `alors poser` $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$, `sinon` $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.
- `Retourner les valeurs` $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T\}$.

Il est important de noter que le calcul de $\alpha(\boldsymbol{X}_t, \boldsymbol{Y})$ dans l'expression (1.10) n'exige pas la connaissance de la constante de normalisation de $\pi(\cdot)$ puisqu'elle disparaît de l'expression à travers le ratio $\pi(\boldsymbol{Y})/\pi(\boldsymbol{X}_t)$. Ainsi, on peut utiliser l'algorithme M-H lorsque la loi à simuler $\pi(\cdot)$ est connue à une constante près, comme par exemple pour les lois *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{y})$ dans un contexte d'inférence bayésienne.

Le noyau de transition $\mathrm{P}(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$ d'une chaîne de Markov construite avec l'algorithme Metropolis-Hastings est

$$\begin{aligned}\mathrm{P}(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t) &= q(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)\alpha(\boldsymbol{X}_t, \boldsymbol{X}_{t+1}) \\ &\quad + \mathbb{1}(\boldsymbol{X}_{t+1} = \boldsymbol{X}_t)[1 - \int q(\boldsymbol{Y}|\boldsymbol{X}_t)\alpha(\boldsymbol{X}_t, \boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}],\end{aligned}$$

où le premier terme résulte de la génération de la variable candidate $\boldsymbol{Y} = \boldsymbol{X}_{t+1}$ selon $q(\cdot|\boldsymbol{X}_t)$ et l'acceptation de celle-ci avec probabilité $\alpha(\boldsymbol{X}_t, \boldsymbol{X}_{t+1})$, tandis que

le deuxième terme résulte du rejet de tous les $\boldsymbol{Y}$ possibles. En utilisant l'expression (1.10), il est facile de vérifier que ce noyau est réversible par rapport à $\pi(\cdot)$, ce qui veut dire

$$\pi(\boldsymbol{X}_t)\,\mathrm{P}(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t) = \pi(\boldsymbol{X}_{t+1})\,\mathrm{P}(\boldsymbol{X}_t|\boldsymbol{X}_{t+1}).$$

Cette propriété est une condition suffisante pour garantir la convergence de la chaîne de Markov construite avec l'algorithme M-H vers $\pi(\cdot)$ pourvu que le choix de la loi de proposition $q(\cdot|\boldsymbol{X}_t)$ assure l'irréductibilité et l'apériodicité de la chaîne (Roberts et Smith, 1994; Tierney, 1994). En termes informels, ceci revient à dire que si x et y sont dans le support de $\pi(\cdot)$, alors il est possible de passer de x à y en un nombre fini d'itérations avec probabilité plus grande que zéro, et que le plus grand commun diviseur de l'ensemble des nombres possibles de sauts effectués pour passer de x à y est égal à 1 (pour tous x, y).

Dans la pratique, les conditions qui garantissent la convergence d'une chaîne de Markov obtenue avec un algorithme M-H sont généralement respectées en prenant une loi de proposition $q(\cdot|\boldsymbol{X}_t)$ qui a le même support que $\pi(\cdot)$. Cependant, le taux de convergence vers la loi limite $\pi(\cdot)$ dépend fortement de la relation entre $q(\cdot|\boldsymbol{X}_t)$ et $\pi(\cdot)$. Le choix d'une loi de proposition inadéquate pour la simulation d'une loi particulière $\pi(\cdot)$ peut entraîner non seulement que la chaîne ait une phase transitoire trop longue, mais aussi qu'elle prenne plus de temps à explorer l'espace dans la phase stationnaire à cause de la forte dépendance des observations (*slow mixing*). D'autre part, la sélection d'une loi de proposition doit tenir compte de l'efficacité computationnelle associée à la génération des réalisations de cette loi et à l'évaluation de l'expression $\alpha(\boldsymbol{X}_t, \boldsymbol{Y})$.

Implémentation

L'implémentation d'un algorithme M-H passe par la spécification d'une loi de proposition. En général, cette loi est choisie entre différentes familles ou typolo-

gies de lois qui demandent souvent l'ajustement de paramètres tels que ceux de dispersion et de position. Pour une classification systématique des typologies de lois disponibles voir Tierney (1994). Dans ce qui suit, on étudie l'algorithme M-H pour l'importante famille de lois symétriques (algorithme Metropolis), ainsi que l'algorithme M-H en blocs, une variante de l'algorithme M-H qui donne lieu au bien connu échantillonneur de Gibbs.

Algorithme Metropolis

L'algorithme Metropolis (Metropolis *et al.*, 1953) est obtenu lorsque la loi de proposition $q(\cdot|\boldsymbol{X})$ est symétrique, c'est-à-dire $q(\boldsymbol{X}|\boldsymbol{Y}) = q(\boldsymbol{Y}|\boldsymbol{X})$ pour tout $\boldsymbol{X}$ et $\boldsymbol{Y}$. Dans ce cas, la probabilité d'acceptation (1.10) devient :

$$\alpha(\boldsymbol{X}, \boldsymbol{Y}) = \min\left(1, \frac{\pi(\boldsymbol{Y})}{\pi(\boldsymbol{X})}\right).$$

Ainsi, si $\pi(\boldsymbol{Y}) \geq \pi(\boldsymbol{X})$ le saut est accepté automatiquement, sinon le saut est accepté avec probabilité égale au rapport des densités.

Un cas spécial de l'algorithme Metropolis survient lorsque la loi de proposition est de la forme $q(\boldsymbol{X}|\boldsymbol{Y}) = q(|\boldsymbol{X} - \boldsymbol{Y}|)$, où $q(\cdot)$ est une loi multivariée. Alors, le candidat $\boldsymbol{Y}$ est généré suivant le processus $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$, où $\boldsymbol{Z}$ est un vecteur aléatoire issu de la loi $q(\cdot)$. Puisque le candidat est égal à l'état actuel de la chaîne plus une perturbation aléatoire $\boldsymbol{Z}$, la chaîne ainsi générée s'appelle *marche aléatoire*. La densité normale multivariée ou celle de la loi de Student multivariée constituent des choix communs pour la loi $q(\cdot)$.

On souligne que, dans le cas d'une marche aléatoire, le paramètre d'échelle de la loi $q(\cdot)$, typiquement une matrice de covariances $\boldsymbol{\Sigma}$, doit être choisi avec précaution afin que les sauts proposés soient à une distance raisonnable dans l'espace. Si

le paramètre d'échelle est tel que les sauts $\boldsymbol{Y} - \boldsymbol{X}_t$ sont trop petits, le taux d'acceptation est élevé mais on explore l'espace trop lentement. Par contre, si l'ajustement de l'échelle produit de trop grands sauts, le taux d'acceptation est trop faible et la chaîne ne change pas assez souvent d'état. Dans les deux cas, on obtient des chaînes ayant des difficultés à bien explorer l'espace d'états.

Algorithme Metropolis-Hastings en blocs

Une alternative à l'actualisation du vecteur $\boldsymbol{X}$ en un seul bloc consiste à diviser $\boldsymbol{X}$ en k composantes ou sous-vecteurs $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k\}$ de dimensions pas nécessairement égales, et actualiser les composantes $\boldsymbol{X}_i$, $i = 1, \ldots, k$, une à la fois. Ainsi, une itération de l'algorithme M-H en blocs est composée de k étapes d'actualisation.

Soit $\boldsymbol{X}_{-i} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_k\}$ et $\boldsymbol{X}_i^t$ l'état de $\boldsymbol{X}_i$ à la fin de l'itération t. À l'étape i de l'itération $t+1$, $\boldsymbol{X}_i$ est actualisé suivant une démarche Metropolis-Hastings. Ainsi, on propose une réalisation candidate $\boldsymbol{Y}_i$ pour $\boldsymbol{X}_i^{t+1}$ avec une loi de proposition $q_i(\boldsymbol{Y}_i|\boldsymbol{X}_i^t, \boldsymbol{X}_{-i}^t)$, où $\boldsymbol{X}_{-i}^t$ désigne la valeur de $\boldsymbol{X}_{-i}$ après avoir complété l'étape $i-1$ de l'itération $t+1$: $\boldsymbol{X}_{-i}^t = \{\boldsymbol{X}_1^{t+1}, \ldots, \boldsymbol{X}_{i-1}^{t+1}, \boldsymbol{X}_{i+1}^t, \ldots, \boldsymbol{X}_k^t\}$.

La probabilité d'acceptation de la réalisation candidate $\boldsymbol{Y}_i$ est :

$$\alpha(\boldsymbol{X}_i^t, \boldsymbol{Y}_i|\boldsymbol{X}_{-i}^t) = \min\left(1, \frac{\pi(\boldsymbol{Y}_i|\boldsymbol{X}_{-i}^t) q_i(\boldsymbol{X}_i^t|\boldsymbol{Y}_i, \boldsymbol{X}_{-i}^t)}{\pi(\boldsymbol{X}_i^t|\boldsymbol{X}_{-i}^t) q_i(\boldsymbol{Y}_i|\boldsymbol{X}_i^t, \boldsymbol{X}_{-i}^t)}\right). \tag{1.11}$$

Si $\boldsymbol{Y}$ est acceptée, alors on pose $\boldsymbol{X}_i^{t+1} = \boldsymbol{Y}_i$, sinon $\boldsymbol{X}_i^{t+1} = \boldsymbol{X}_i^t$.

Les lois $\pi(\boldsymbol{X}_i|\boldsymbol{X}_{-i})$, $i = 1, \ldots, k$, sont nommées les lois conditionnelles complètes (*full conditional distributions*). La loi conditionnelle complète de la composante $\boldsymbol{X}_i$ est

$$\pi(\boldsymbol{X}_i|\boldsymbol{X}_{-i}) = \frac{\pi(\boldsymbol{X})}{\int \pi(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X}_i}.$$

Un cas spécial très important de l'algorithme M-H en blocs est l'*échantillonneur de Gibbs* (Geman et Geman, 1984; Gelfand et Smith, 1990). Dans l'échantillonneur de Gibbs, les lois de proposition $q_i(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{X}_{-i})$ utilisées pour actualiser les composantes $\boldsymbol{X}_i$ sont les lois conditionnelles complètes $\pi(\boldsymbol{X}_i|\boldsymbol{X}_{-i}), i = 1, \ldots, k$. À partir de l'expression (1.11) on déduit que, dans ce cas, la probabilité d'acceptation est égale à 1 et que les réalisations candidates $\boldsymbol{Y}_i$ sont toujours acceptées. Ainsi, l'échantillonneur de Gibbs consiste simplement à générer des réalisations des lois conditionnelles complètes itérativement.

On souligne que, dans un algorithme M-H en blocs, rien ne nous empêche de combiner des étapes Gibbs, où l'on actualise certaines composantes de $\boldsymbol{X}$ directement à partir des lois conditionnelles complètes, avec des étapes où l'on utilise la forme plus générale de l'algorithme M-H. Ceci est souvent le cas lorsqu'on a affaire à des lois conditionnelles complètes pour lesquelles la génération directe d'échantillons n'est pas possible ou est très coûteuse du point de vue computationnel. L'algorithme utilisé pour l'analyse multivariée bayésienne des données dans le cadre de notre étude, présenté dans les prochaines sections, est un exemple d'implémentation de cette forme particulière de l'algorithme M-H.

1.4 Spécification bayésienne du modèle et approximation *t-link*

Dans la section 1.2.3, on a présenté la loi logistique multivariée proposée par O'Brien et Dunson (2004) ainsi que la vraisemblance (1.5) construite à partir de cette dernière pour l'analyse de données binaires par régression logistique multivariée. Afin d'adopter une approche bayésienne pour l'inférence sur les paramètres $\boldsymbol{\beta}$ et $\boldsymbol{R}$ du modèle, il est nécessaire de définir une loi *a priori* $p(\boldsymbol{\beta}, \boldsymbol{R})$. En suivant le choix fait par O'Brien et Dunson (2004), on suppose, par simplicité, l'indépendance *a priori* de $\boldsymbol{\beta}$ et $\boldsymbol{R}$, c'est-à-dire $p(\boldsymbol{\beta}, \boldsymbol{R}) = p(\boldsymbol{\beta})p(\boldsymbol{R})$. La loi *a priori* $p(\boldsymbol{\beta})$ assignée aux coefficients $\boldsymbol{\beta}$ de la régression est une loi normale multivariée

$\mathcal{N}_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. La loi $p(\boldsymbol{R})$ pour les $p(p-1)/2$ coefficients qui déterminent de façon unique la matrice de corrélation $\boldsymbol{R}$ est une loi uniforme avec support sur l'espace de matrices de ce type.

Ainsi, la loi *a posteriori* des paramètres prend la forme :

$$p(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{R}) p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R}),$$

où $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R}) = \prod_{i=1}^{n} p(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{R})$ et $p(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{R})$ est la vraisemblance associée au $i^{\text{ème}}$ survivant. Or, la complexité de la vraisemblance $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R})$ rend difficile la simulation efficace de la loi $p(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$. Une solution à ce problème consiste à approcher la vraisemblance $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R})$ par une vraisemblance alternative $p^*(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R})$, de façon telle que la simulation de la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{R}) p^*(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R})$ puisse se faire relativement facilement à l'aide d'un algorithme MCMC. Si l'approximation est assez bonne, on peut faire l'inférence sur les paramètres $\boldsymbol{\beta}$ et $\boldsymbol{R}$ du modèle directement à partir de la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$. En outre, la méthode d'échantillonnage d'importance (Hastings, 1970) peut nous permettre d'estimer exactement les caractéristiques de la loi *a posteriori* $p(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$ en pondérant de façon appropriée les réalisations de la loi $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$.

C'est dans cet esprit que O'Brien et Dunson (2004), inspirés par les résultats d'Albert et Chib (1993) montrant la grande similitude entre la loi logistique et la loi de Student univariée lorsque $\nu = 8$, ont proposé d'approcher la densité de la loi logistique multivariée $\mathcal{L}_p(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{R}, \nu)$ de la vraisemblance (1.5) par celle d'une loi de Student multivariée $\mathcal{T}_p(\boldsymbol{z}|\boldsymbol{\mu}, \sigma^2\boldsymbol{R}, \nu)$. Afin de rendre l'approximation presque exacte, ils ont suggéré de fixer σ^2 à $\pi^2(\nu-2)/3\nu$ pour égaliser les variances de chaque loi, et de fixer le paramètre ν de la loi logistique à 7.3 pour minimiser l'erreur quadratique intégrée entre les densités logistique et Student univariées. L'application de cette approximation pour l'analyse multivariée des données issues

d'une étude de neurotoxicité (O'Brien et Dunson, 2004) a permis de constater que les résultats étaient très peu influencés par le fait de pondérer ou non les échantillons obtenus de la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$ par des poids d'importance. Par conséquent, on a décidé d'adopter cette même approximation (modèle *t-link*) pour notre étude. Ainsi, la nouvelle vraisemblance pour les vecteurs de réponses métaboliques est :

$$\mathrm{P}(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{R}) \approx$$
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{j=1}^{p} \mathbb{1}(z_{ij} > 0)^{y_{ij}} \mathbb{1}(z_{ij} \leq 0)^{1-y_{ij}} \right) \mathcal{T}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}, \nu)\, \mathrm{d}\boldsymbol{z}_i. \quad (1.12)$$

Bien que la complexité des calculs pour l'évaluation de cette vraisemblance rend très difficile l'usage direct de la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$ pour l'inférence bayésienne, le modèle t-link a l'avantage d'admettre une démarche alternative qui permet d'améliorer considérablement l'efficacité des calculs pour l'échantillonnage de la loi *a posteriori*. Dans cette nouvelle approche, au lieu de travailler avec la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}|\boldsymbol{y})$, on s'intéresse à la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{y})$ où $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_n)$ est un vecteur de variables latentes spécifiées subséquemment. L'obtention de la loi $p^*(\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{y})$ passe par la construction d'une vraisemblance qui fait intervenir les vecteurs $\boldsymbol{Z}$ et $\boldsymbol{\Phi}$. En effet, on introduit d'abord le vecteur aléatoire $\boldsymbol{Z}_i$ dans la vraisemblance associée au $i^{\text{ème}}$ survivant. Ainsi, puisque :

$$Y_{ij} = \mathbb{1}(Z_{ij} > 0) \quad j = 1, \ldots, p$$
$$\boldsymbol{Z}_i|\boldsymbol{\beta}, \boldsymbol{R} \sim \mathcal{T}_p(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}, \nu)$$

on a :

$$p^*(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}) = p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}) p^*(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{z}_i)$$

où $p^*(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{z}_i) = \prod_{j=1}^{p} (\mathbb{1}(z_{ij} > 0)\mathbb{1}(y_{ij} = 1) + \mathbb{1}(z_{ij} \leq 0)\mathbb{1}(y_{ij} = 0))$ et $p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}) = \mathcal{T}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}, \nu)$.

D'autre part, par la définition 1.2.1, on sait que le vecteur $\boldsymbol{Z}_i$ peut être exprimé comme :

$$\boldsymbol{Z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \frac{1}{\sqrt{\Phi_i}}\boldsymbol{W},$$

où Φ_i est une variable aléatoire distribuée selon une loi gamma $\text{Gam}(\nu/2, \nu/2)$ et $\boldsymbol{W}$ est un vecteur aléatoire normalement distribué $\mathcal{N}(0, \sigma^2\boldsymbol{R})$ indépendant de Φ_i.

Alors, on déduit que lorsque $\Phi_i = \phi_i$, on a $\boldsymbol{Z}_i|\phi_i \sim \mathcal{N}_p(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\phi_i^{-1}\boldsymbol{R})$. Ce résultat nous permet d'introduire la deuxième variable latente Φ_i dans le modèle :

$$\begin{aligned} Y_{ij} &= \mathbb{1}(Z_{ij} > 0) \quad j = 1, \ldots, p \\ \boldsymbol{Z}_i|\boldsymbol{\beta}, \boldsymbol{R}, \Phi_i &\sim \mathcal{N}_p(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\phi_i^{-1}\boldsymbol{R}) \\ \Phi_i|\boldsymbol{\beta}, \boldsymbol{R} &\sim \text{Gam}(\nu/2, \nu/2). \end{aligned}$$

Ainsi, on obtient :

$$p^*(\boldsymbol{y}_i, \phi_i, \boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}) = p^*(\phi_i|\boldsymbol{\beta}, \boldsymbol{R})p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i)p^*(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i, \boldsymbol{z}_i),$$

où $p^*(\phi_i|\boldsymbol{\beta}, \boldsymbol{R}) = \text{Gam}(\phi_i|\nu/2, \nu/2)$, $p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i) = \mathcal{N}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\phi_i^{-1}\boldsymbol{R})$ et $p^*(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i, \boldsymbol{z}_i) \equiv p^*(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{z}_i)$.

Finalement, la loi *a posteriori* $p^*(\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{y})$ prend la forme :

$$\begin{aligned} p^*(\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{y}) &\propto p(\boldsymbol{\beta}, \boldsymbol{R})p^*(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{R}) \\ &\propto p(\boldsymbol{\beta}, \boldsymbol{R})p^*(\boldsymbol{\phi}|\boldsymbol{\beta}, \boldsymbol{R})p^*(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi})p^*(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{z}), \quad (1.13) \end{aligned}$$

où $p^*(\boldsymbol{\phi}|\boldsymbol{\beta},\boldsymbol{R}) = \prod_{i=1}^{n} p^*(\phi_i|\boldsymbol{\beta},\boldsymbol{R})$, $p^*(\boldsymbol{z}|\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi}) = \prod_{i=1}^{n} p^*(\boldsymbol{z}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i)$ et
$p^*(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z}) = \prod_{i=1}^{n} p^*(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i,\boldsymbol{z}_i)$.

Dans la section suivante, on décrit l'algorithme MCMC proposé par O'Brien et Dunson (2004) ; c'est cet algorithme qu'on a implémenté pour générer des réalisations de la loi $p^*(\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{y})$ dans le contexte de notre étude.

1.5 Algorithme MCMC

L'algorithme MCMC pour échantillonner la loi $p^*(\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{y})$ est un algorithme M-H en blocs : à chaque itération on actualise de manière séquentielle les blocs : $\boldsymbol{Z}_i,\phi_i, i = 1,\ldots,n$, $\boldsymbol{\beta}$ et $\boldsymbol{R}$. Lorsque $p(\boldsymbol{\beta},\boldsymbol{R}) = p(\boldsymbol{\beta})p(\boldsymbol{R})$ avec $\boldsymbol{\beta} \sim \mathcal{N}_q(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, les lois conditionnelles complètes de tous les blocs, à l'exception de $\boldsymbol{R}$, prennent des formes standards. Ainsi, l'algorithme alterne des étapes Gibbs pour l'actualisation des composantes $\boldsymbol{Z}_i,\phi_i$ $(i = 1,\ldots,n)$ et $\boldsymbol{\beta}$ avec une étape Metropolis plus générale pour le bloc $\boldsymbol{R}$.

Maintenant, on dérive les lois conditionnelles complètes pour les blocs : $\boldsymbol{Z}_i,\phi_i, i = 1,\ldots,n$, $\boldsymbol{\beta}$ et $\boldsymbol{R}$ à partir de l'expression de la loi $p^*(\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{y})$. D'abord, on remarque que la composante $\boldsymbol{z}_i$ apparaît dans les termes $p^*(\boldsymbol{z}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i)$ et $p^*(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i,\boldsymbol{z}_i)$ de la loi *a posteriori*. Ainsi, on déduit :

$$\begin{aligned} p^*(\boldsymbol{z}_i|\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{y},\boldsymbol{z}_{-i}) &\propto p^*(\boldsymbol{z}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i)p^*(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i,\boldsymbol{z}_i) \\ &\propto \mathcal{N}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\sigma^2\phi_i^{-1}\boldsymbol{R}) \\ &\quad \prod_{j=1}^{p} (\mathbb{1}(z_{ij} > 0)\mathbb{1}(y_{ij} = 1) + \mathbb{1}(z_{ij} \leq 0)\mathbb{1}(y_{ij} = 0)) \quad (1.14) \end{aligned}$$

et donc $\boldsymbol{Z}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i,\boldsymbol{y}_i \sim TN_p(\boldsymbol{X}_i\boldsymbol{\beta},\sigma^2\phi_i^{-1}\boldsymbol{R},\Omega_y)$, où $\Omega_y = \{\boldsymbol{z}_i \in \mathbb{R}^p : z_{ij} > 0$ si $y_{ij} = 1, z_{ij} \leq 0$ sinon, $j = 1,\ldots,p\}$, dénote le support de la loi normale

tronquée multivariée.

En appliquant la même démarche pour les variables latentes ϕ_i, $i = 1, \ldots, n$, on obtient :

$$\begin{aligned} p^*(\phi_i|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\phi}_{-i}, \boldsymbol{y}, \boldsymbol{z}) &\propto p^*(\phi_i|\boldsymbol{\beta}, \boldsymbol{R}) p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i) \\ &\propto \text{Gam}(\phi_i|\nu/2, \nu/2) \mathcal{N}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\phi_i^{-1}\boldsymbol{R}) \\ &\propto \phi_i^{(\nu+p)/2-1} \\ &\quad \exp\left(-\frac{\nu + \sigma^{-2}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{R}^{-1}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})}{2}\phi_i\right), \end{aligned}$$

ce qui implique

$$\Phi_i|\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{z}_i \sim \text{Gam}\left(\frac{\nu+p}{2}, \frac{\nu + \sigma^{-2}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{R}^{-1}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})}{2}\right).$$

Pour le vecteur $\boldsymbol{\beta}$ on a :

$$\begin{aligned} p^*(\boldsymbol{\beta}|\boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{z}) &\propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} p^*(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{R}, \phi_i) \\ &\propto \mathcal{N}_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \prod_{i=1}^{n} \mathcal{N}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\phi_i^{-1}\boldsymbol{R}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right) \\ &\quad \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\sigma^{-2}\phi_i\boldsymbol{R}^{-1}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right). \quad (1.15) \end{aligned}$$

Ensuite, après regroupement des termes en $\boldsymbol{\beta}$, l'expression (1.15) devient :

$$\begin{aligned} p^*(\boldsymbol{\beta}|\boldsymbol{R}, \boldsymbol{\phi}, \boldsymbol{y}, \boldsymbol{z}) \propto \exp\Big(&-\frac{1}{2}\big(\boldsymbol{\beta}'(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sigma^{-2}\sum_{i=1}^{n}\phi_i\boldsymbol{X}_i'\boldsymbol{R}^{-1}\boldsymbol{X}_i)\boldsymbol{\beta} \\ &-2\boldsymbol{\beta}'(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^{n}\phi_i\boldsymbol{X}_i'\boldsymbol{R}^{-1}\boldsymbol{z}_i)\big)\Big). \end{aligned}$$

En notant $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sigma^{-2}\sum_{i=1}^{n}\phi_i\boldsymbol{X}_i'\boldsymbol{R}^{-1}\boldsymbol{X}_i)^{-1}$ et $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 +$ $+\sigma^{-2}\sum_{i=1}^{n}\phi_i\boldsymbol{X}_i'\boldsymbol{R}^{-1}\boldsymbol{z}_i)$, on obtient :

$$p^*(\boldsymbol{\beta}|\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{y},\boldsymbol{z}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}})'\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}})\right).$$

et donc $\boldsymbol{\beta}|\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z} \sim \mathcal{N}_q(\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}},\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$.

Finalement, pour les blocs $\boldsymbol{R}$ on a :

$$\begin{aligned} p^*(\boldsymbol{R}|\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{y},\boldsymbol{z}) &\propto p(\boldsymbol{R})\prod_{i=1}^{n} p^*(\boldsymbol{z}_i|\boldsymbol{\beta},\boldsymbol{R},\phi_i) \\ &\propto p(\boldsymbol{R})\prod_{i=1}^{n} \mathcal{N}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\sigma^2\phi_i^{-1}\boldsymbol{R}) \\ &\propto p(\boldsymbol{R})|\boldsymbol{R}|^{-n/2}\exp\left(-\frac{\sigma^{-2}}{2}\sum_{i=1}^{n}\phi_i(\boldsymbol{z}_i-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{R}^{-1}(\boldsymbol{z}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right). \end{aligned}$$

On remarque que, indépendamment du choix pour $p(\boldsymbol{R})$, la loi conditionnelle complète associée au bloc $\boldsymbol{R}$ n'est pas standard et qu'il n'est pas possible de générer des réalisations de $\boldsymbol{R}$ avec des méthodes de simulation directe. On souligne aussi que cette dernière loi est la loi associée aux $p(p-1)/2$ coefficients qui déterminent de manière univoque la matrice de corrélation $\boldsymbol{R}$.

Maintenant, on présente le schéma algorithmique utilisé pour la simulation de la loi *a posteriori* $p^*(\boldsymbol{\beta},\boldsymbol{R},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{y})$, avec un échantillon résultant pour la loi a posteriori marginale de $\boldsymbol{\beta}$ et $\boldsymbol{R}$:

— `Initialiser` $\boldsymbol{Z}_i^{(0)},\Phi_i^{(0)},i=1,\ldots,n,\boldsymbol{\beta}^{(0)}$ et $\boldsymbol{R}^{(0)}$.

— `Pour` $t=0,1,2,\ldots,T-1$:

— `Pour` $i=1,\ldots,n$, `générer` $\boldsymbol{Z}_i^{(t+1)}$ `de la loi conditionnelle complète de` $\boldsymbol{Z}_i$:

$$\boldsymbol{Z}_i^{(t+1)} \sim TN_p\big(\boldsymbol{X}_i\boldsymbol{\beta}^{(t)},\sigma^2\big(\phi_i^{(t)}\big)^{-1}\boldsymbol{R}^{(t)},\Omega_y\big).$$

— Pour $i = 1, \ldots, n$, générer $\Phi_i^{(t+1)}$ de la loi conditionnelle complète de Φ_i :

$$\Phi_i^{(t+1)} \sim \text{Gam}\left(\frac{\nu+p}{2}, \frac{\nu+\sigma^{-2}\left(\boldsymbol{z}_i^{(t+1)} - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}\right)'\left(\boldsymbol{R}^{(t)}\right)^{-1}\left(\boldsymbol{z}_i^{(t+1)} - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}\right)}{2}\right).$$

— Générer $\boldsymbol{\beta}^{(t+1)}$ de la loi conditionnelle complète: $\boldsymbol{\beta} \sim \mathcal{N}_q\left(\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right)$, où

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sigma^{-2}\sum_{i=1}^{n}\phi_i^{(t+1)}\boldsymbol{X}_i'\left(\boldsymbol{R}^{(t)}\right)^{-1}\boldsymbol{X}_i\right)^{-1}$$

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^{n}\phi_i^{(t+1)}\boldsymbol{X}_i'\left(\boldsymbol{R}^{(t)}\right)^{-1}\boldsymbol{z}_i^{(t+1)}\right).$$

— Générer une réalisation candidate $\boldsymbol{R}_{\text{prop}}$ pour les $p^* = p(p-1)/2$ éléments uniques de la matrice $\boldsymbol{R}$:

$$\text{unique}\,\boldsymbol{R}_{\text{prop}} \sim \mathcal{N}_{p^*}(\text{unique}\,\boldsymbol{R}^{(t)}, \boldsymbol{\Omega}).$$

— Générer U de la loi $\mathcal{U}(0,1)$.

— Si

$$U \leq \min\left(1, \frac{p\left(\boldsymbol{R}_{\text{prop}}\right)\prod_{i=1}^{n}\mathcal{N}_p(\boldsymbol{z}_i^{(t+1)}|\boldsymbol{X}_i\boldsymbol{\beta}^{(t+1)}, \sigma^2\left(\phi_i^{(t+1)}\right)^{-1}\boldsymbol{R}_{\text{prop}})}{p\left(\boldsymbol{R}^{(t)}\right)\prod_{i=1}^{n}\mathcal{N}_p(\boldsymbol{z}_i^{(t+1)}|\boldsymbol{X}_i\boldsymbol{\beta}^{(t+1)}, \sigma^2\left(\phi_i^{(t+1)}\right)^{-1}\boldsymbol{R}^{(t)})}\right)$$

alors poser $\boldsymbol{R}^{(t+1)} = \boldsymbol{R}_{\text{prop}}$, sinon $\boldsymbol{R}^{(t+1)} = \boldsymbol{R}^{(t)}$.

— Retourner les valeurs $\{(\boldsymbol{\beta}^{(1)}, \boldsymbol{R}^{(1)}), (\boldsymbol{\beta}^{(2)}, \boldsymbol{R}^{(2)}), \ldots, (\boldsymbol{\beta}^{(T)}, \boldsymbol{R}^{(T)})\}$.

1.6 Simulation de la loi normale tronquée multivariée

L'algorithme MCMC présenté antérieurement pour la simulation de la loi *a posteriori* (1.13) doit être capable de générer des réalisations pour différentes lois de façon itérative. En particulier, on a vu que l'actualisation de chaque bloc $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, demande la simulation de lois normales tronquées multivariées (NMVT). Or, il n'existe pas de méthode simple pour simuler de cette famille de lois efficacement. L'échantillonneur de Gibbs est la technique la plus répandue dans cette situation puisque les lois conditionnelles complètes d'une loi NMVT sont des lois normales tronquées univariées (NUVT) pour lesquelles il est possible de générer des réalisations directement. Par conséquent, on a décidé d'effectuer les actualisations des blocs $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, dans notre algorithme MCMC principal au moyen d'un échantillonneur de Gibbs.

Dans ce qui suit, on dérive les lois conditionnelles complètes d'une loi NMVT. De plus, on présente le schéma de simulation utilisée ainsi que son intégration dans notre algorithme MCMC principal.

1.6.1 La loi normale tronquée multivariée et les lois conditionnelles complètes

Définition 1.6.1. Un vecteur aléatoire $\boldsymbol{W}$ de dimension p est distribué selon une loi normale tronquée multivariée (NMVT) dont le support est l'hyper-rectangle $\mathcal{R} = \prod_{j=1}^{p} \mathcal{R}_j$ ($\mathcal{R}_j = (a_j, b_j)$) si sa densité est de la forme :

$$p(\boldsymbol{w}) = \frac{\exp(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu}))}{\int_{\mathcal{R}} \exp(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu}))\,\mathrm{d}\boldsymbol{w}} \mathbb{1}(\boldsymbol{w} \in \mathcal{R}), \tag{1.16}$$

où $\boldsymbol{\Sigma}$ est une matrice définie positive. On note $\boldsymbol{W} \sim TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{R})$.

On souligne que les paramètres a_j et b_j, $j = 1, \ldots, p$, apparaissant dans le support $\mathcal{R}$ peuvent être $-\infty$ et $+\infty$. Dans notre cas, les lois conditionnelles complètes des

blocs $\boldsymbol{Z}_i$ sont en fait des lois NMVT avec $\mathcal{R}_{ij} = (0, +\infty)$ ou $\mathcal{R}_{ij} = (-\infty, 0)$, $j = 1, \ldots, p$, selon les valeurs prises par les variables binaires Y_{ij}. En outre, bien que dans cette section on se concentre sur la loi NMVT de la définition 1.6.1, il est possible de définir une famille plus large de lois NMVT en considérant comme support une région convexe quelconque de $\mathbb{R}^p$.

Maintenant, on dérive les lois conditionnelles complètes de la loi NMVT. D'abord, on réécrit la densité (1.16) au moyen de la densité normale multivariée $\mathcal{N}_p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\boldsymbol{w}) = \frac{\mathcal{N}_p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\mathcal{R}} \mathcal{N}_p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{w}} \mathbb{1}(\boldsymbol{w} \in \mathcal{R}). \tag{1.17}$$

En utilisant (1.17), les densités $p(w_j|\boldsymbol{w}_{-j})$, avec $\boldsymbol{w}_{-j} \in \prod_{\substack{k=1 \\ k \neq j}}^{p} \mathcal{R}_k$, et $j = 1, \ldots, p$, prennent la forme :

$$\begin{aligned} p(w_j|\boldsymbol{w}_{-j}) = \frac{p(\boldsymbol{w})}{p(\boldsymbol{w}_{-j})} = \frac{p(\boldsymbol{w})}{\int_{a_j}^{b_j} p(\boldsymbol{w})\, \mathrm{d}w_j} &= \frac{\mathcal{N}_p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{a_j}^{b_j} \mathcal{N}_p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}w_j} \mathbb{1}(w_j \in (a_j, b_j)) \\ &\propto \mathcal{N}_p((w_j, \boldsymbol{w}_{-j})|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathbb{1}(w_j \in (a_j, b_j)). \end{aligned}$$

D'autre part, on sait que si $\boldsymbol{X} = (X_j, \boldsymbol{X}_{-j}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ où $\boldsymbol{\mu} = (\mu_j, \boldsymbol{\mu}_{-j})'$ et $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{jj}^2 & \boldsymbol{\Sigma}_{j,-j} \\ \boldsymbol{\Sigma}_{-j,j} & \boldsymbol{\Sigma}_{-j,-j} \end{pmatrix}$, avec $\boldsymbol{\Sigma}_{-j,-j}$ définie positive, alors :

$$p(x_j|\boldsymbol{x}_{-j}) = \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}_{-j})} = \frac{\mathcal{N}_p((x_j, \boldsymbol{x}_{-j})|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\mathbb{R}} \mathcal{N}_p((x_j, \boldsymbol{x}_{-j})|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}x_j} = \mathcal{N}(x_j|\mu_{j.-j}, \sigma_{j.-j}^2),$$

où $\sigma_{j.-j}^2 = \sigma_{jj}^2 - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$ et $\mu_{j.-j} = \mu_j + \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}(\boldsymbol{x}_{-j} - \boldsymbol{\mu}_{-j})$.

Ainsi, on déduit que les lois conditionnelles complètes de la loi NMVT sont des lois normales tronquées univariées (NUVT) dont les densités sont de la forme :

$$p(w_j|\boldsymbol{w}_{-j}) \propto \mathcal{N}_p(w_j|\mu_{j.-j}, \sigma_{j.-j}^2) \mathbb{1}(w_j \in (a_j, b_j)),\ j = 1, \ldots, p,$$

c'est-à-dire : $W_j|\boldsymbol{W}_{-j} \sim TN_1(\mu_{j.-j}, \sigma_{j.-j}^2, \mathcal{R}_j)$.

1.6.2 Simulation de la loi NUVT et échantillonneur de Gibbs

La simulation de la loi NMVT au moyen d'un échantillonneur de Gibbs demande de pouvoir simuler des lois NUVT. À cette fin, et dans un premier temps, on a choisi la méthode d'inversion due à sa simplicité.

La fonction de répartition d'une variable aléatoire $W \sim TN_1(\mu, \sigma^2, (a, b))$ est :

$$F(w) = \begin{cases} 0 & \text{si } w < a \\ \dfrac{\Phi(\frac{w-\mu}{\sigma}) - p_1}{p_2 - p_1} & \text{si } a \leq w \leq b, \\ 1 & \text{si } w > b \end{cases}$$

où Φ est la fonction de répartition d'une variable normale centrée réduite, $p_1 = \Phi(\frac{a-\mu}{\sigma})$ et $p_2 = \Phi(\frac{b-\mu}{\sigma})$.

Pour générer une réalisation de cette loi avec la méthode d'inversion, on doit résoudre l'équation $F(W) = U$, où $U \sim \mathcal{U}(0, 1)$. Après une simple manipulation algébrique on obtient :

$$W = \mu + \sigma\Phi^{-1}(p_1 + U(p_2 - p_1)).$$

Ainsi, l'échantillonneur de Gibbs pour la simulation d'un vecteur aléatoire $\boldsymbol{W} \sim TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{R})$, où $\mathcal{R} = \prod_{j=1}^{p}(a_j, b_j)$, $(a_j, b_j) = (0, +\infty)$ ou $(-\infty, 0)$, prend la forme suivante :

- `Initialiser` $\boldsymbol{W}^{(0)} = (W_1^{(0)}, W_2^{(0)}, \ldots, W_p^{(0)})$.
- `Précalculer` $\boldsymbol{\Sigma}_{-j,-j}^{-1}$ `et` $\sigma_{j.-j}^2 = \sigma_{jj}^2 - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$, $j = 1, \ldots, p$.
- `Pour` $t = 0, 1, \ldots, T-1$:
  - `Pour` $j = 1, \ldots, p$ `générer` $W_j^{(t+1)}$ `de la loi conditionnelle complète` $TN_1(\mu_{j.-j}^{(t)}, \sigma_{j.-j}^2, (a_j, b_j))$ où $\mu_{j.-j}^{(t)} = \mu_j + \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}(\boldsymbol{W}_{-j}^{(t)} - \boldsymbol{\mu}_{-j})$:

— `Générer` U de la loi $\mathcal{U}(0,1)$.

— `Si` $a_j = 0$ $(\mathcal{R}_j = (0, +\infty))$:

$$W_j^{(t+1)} = \mu_{j.-j}^{(t)} + \sigma_{j.-j}\Phi^{-1}\left(\Phi\left(\frac{-\mu_{j.-j}^{(t)}}{\sigma_{j.-j}}\right) + U\left(1 - \Phi\left(\frac{-\mu_{j.-j}^{(t)}}{\sigma_{j.-j}}\right)\right)\right),$$

`sinon` $(\mathcal{R}_j = (-\infty, 0))$:

$$W_j^{(t+1)} = \mu_{j.-j}^{(t)} + \sigma_{j.-j}\Phi^{-1}\left(U\left(\Phi\left(\frac{-\mu_{j.-j}^{(t)}}{\sigma_{j.-j}}\right)\right)\right).$$

— `Retourner les valeurs` $\{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(T)}\}$.

Dans un deuxième temps, afin d'accroître la précision des calculs dans les simulations des lois NUVT, on a décidé d'adopter l'algorithme d'acceptation-rejet mixte proposé par Li et Ghosh (1993). Cet algorithme, qui est une version améliorée des méthodes de simulation développées par Geweke (1991) et Robert (1995), est basé sur une utilisation intelligente de quatre méthodes de rejet standards dans le but d'optimiser le taux d'acceptation global des propositions. D'abord, l'algorithme tire profit d'une propriété fondamentale de la loi NUVT qu'on présente dans la proposition 1.6.1.

Proposition 1.6.1. Soit $W \sim TN_1(\mu, \sigma^2, \mathcal{R}^1)$, $\mathcal{R}^1 = (c, d)$. Alors

$$X = \frac{W - \mu}{\sigma} \sim TN_1(0, 1, \mathcal{R}^2),$$

où $\mathcal{R}^2 = (a, b)$ avec $a = (c - \mu)/\sigma$ et $b = (d - \mu)/\sigma$.

Ainsi, une fois qu'on a généré une réalisation x de la variable aléatoire $X \sim TN_1(0, 1, \mathcal{R}^2)$, on peut facilement obtenir une réalisation de $W \sim TN_1(\mu, \sigma^2, \mathcal{R}^1)$ au moyen de la transformation $w = \mu + \sigma x$. En conséquence, l'algorithme de Li et Ghosh présuppose, sans perte de généralité, que la variable aléatoire à simuler est toujours de loi NUVT standard.

On remarque que lorsque $\mathcal{R}^1 = (0, +\infty)$, $\mathcal{R}^2$ est de la forme $(a, +\infty)$. Dans ce cas particulier, l'algorithme propose trois lois de proposition possibles en fonction de la valeur de a :

i. $a < 0$: on utilise une loi normale standard ;

ii. $0 < a < a_0 = 0.2570$: on prend une loi normale standard tronquée à la moyenne (0) ;

iii. $a \geq a_0 = 0.2570$: la loi de proposition est une loi exponentielle de paramètre $\lambda^* = \frac{a+\sqrt{a^2+4}}{2}$ translatée de a unités.

Les valeurs de a_0 et λ^* sont telles que le taux d'acceptation de l'algorithme qui combine ces trois lois est maximisé. Concernant la loi de proposition exponentielle translatée, il est possible de montrer que le taux d'acceptation optimisé est égal à $\sqrt{2\pi}\lambda^* \exp(-\frac{\lambda^{*2}}{2} + \lambda^* a)\Phi(-a)$. On souligne aussi que pour des intervalles de la forme $\mathcal{R}^2 = (-\infty, a)$ $(\mathcal{R}^1 = (-\infty, 0))$, on peut obtenir une réalisation en appliquant la même méthode sur l'intervalle $\mathcal{R}^2 = (-a, +\infty)$ et en changeant le signe du résultat.

L'application de cet algorithme pour la simulation de lois NUVT avec des intervalles de troncature de la forme $(0, +\infty)$ ou $(-\infty, 0)$ est présentée directement dans le schéma algorithmique de l'échantillonneur de Gibbs suivant :

— `Initialiser` $\boldsymbol{W}^{(0)} = (W_1^{(0)}, W_2^{(0)}, \ldots, W_p^{(0)})$.

— `Précalculer` $\boldsymbol{\Sigma}_{-j,-j}^{-1}$ `et` $\sigma_{j.-j}^2 = \sigma_{jj}^2 - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$, $j = 1, \ldots, p$.

— `Pour` $t = 0, 1, \ldots, T-1$:

— `Pour` $j = 1, \ldots, p$ `générer` $W_j^{(t+1)}$ `de la loi conditionnelle complète` $TN_1(\mu_{j.-j}^{(t)}, \sigma_{j.-j}^2, (a_j, b_j))$ où $\mu_{j.-j}^{(t)} = \mu_j + \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}(\boldsymbol{W}_{-j}^{(t)} - \boldsymbol{\mu}_{-j})$:

— `Si` $a_j = 0$ $(\mathcal{R}_j = (0, +\infty))$: $a = \frac{-\mu_{j.-j}^{(t)}}{\sigma_{j.-j}}$, $s = 1$,

sinon : $a = \frac{\mu_{j.-j}^{(t)}}{\sigma_{j.-j}}$, $s = -1$.

— Si $a < 0$:

1. Générer $Z \sim \mathcal{N}(0,1)$,
2. Si $Z \geq a$: $X = Z$, sinon : retour en 1.

sinon :

— Si $a < 0.2570$:

1. Générer $Z \sim \mathcal{N}(0,1)$,
2. Si $|Z| \geq a$: $X = |Z|$, sinon : retour en 1.

sinon :

1. Calculer $\lambda^* = \dfrac{a + \sqrt{a^2+4}}{2}$,
2. Générer $Z' \sim \text{Exp}(\lambda^*), U \sim \mathcal{U}(0,1)$ et $Z = a + Z'$,
3. Si $U \leq \exp(-1/2(Z - \lambda^*)^2)$: $X = Z$, sinon : retour en 2.

— Poser $W_j^{(t+1)} = \mu_{j.-j}^{(t)} + s\sigma_{j.-j}X$.

— Retourner les valeurs $\{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(T)}\}$.

Quelque soit la méthode utilisée pour la génération des lois NUVT, on souligne que l'échantillonneur de Gibbs peut être formulé en termes de la matrice de précision $\boldsymbol{H} = \boldsymbol{\Sigma}^{-1}$ au lieu de la matrice de covariances $\boldsymbol{\Sigma}$. En effet, si on écrit la matrice de précision $\boldsymbol{H}$ en forme de matrice par blocs :

$$\boldsymbol{H} = \begin{pmatrix} \sigma_{jj}^2 & \boldsymbol{\Sigma}_{j,-j} \\ \boldsymbol{\Sigma}_{-j,j} & \boldsymbol{\Sigma}_{-j,-j} \end{pmatrix}^{-1} = \begin{pmatrix} H_{jj} & \boldsymbol{H}_{j,-j} \\ \boldsymbol{H}_{-j,j} & \boldsymbol{H}_{-j,-j} \end{pmatrix},$$

alors, en utilisant des formules d'inversion, on obtient :

$$\sigma^2_{j.-j} = \sigma^2_{jj} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}^{-1}_{-j,-j}\boldsymbol{\Sigma}_{-j,j} = H^{-1}_{jj}$$
$$\mu_{j.-j} = \mu_j + \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}^{-1}_{-j,-j}(\boldsymbol{W}_{-j} - \boldsymbol{\mu}_{-j}) = \mu_j - H^{-1}_{jj}\boldsymbol{H}_{j,-j}(\boldsymbol{W}_{-j} - \boldsymbol{\mu}_{-j}).$$

On remarque aussi que, lorsque la matrice $\boldsymbol{H}$ est connue, l'échantillonneur de Gibbs ne demande que l'inversion de H_{jj}, qui est un scalaire. Ainsi, l'utilisation de la matrice $\boldsymbol{H}$ au détriment de la matrice $\boldsymbol{\Sigma}$ est souvent préférable en termes d'efficacité computationnelle.

Bien que les actualisations des blocs $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, dans l'algorithme MCMC de la section 1.5 peuvent être faites au moyen de l'un ou l'autre échantillonneur de Gibbs proposés ici, l'inclusion directe de cet échantillonneur dans un algorithme Metropolis-Hastings en blocs est trop lourde computationnellement. Une alternative qui s'est avérée plus efficace consiste à traiter chaque composante Z_{ij}, $j = 1, \ldots, p$, de chaque vecteur $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, comme un bloc en lui même. À partir de l'expression (1.14) des densités des lois conditionnelles complètes de chaque vecteur $\boldsymbol{Z}_i$, il est facile de vérifier que les lois conditionnelles complètes des blocs Z_{ij} $(j = 1, \ldots, p;\ i = 1, \ldots, n)$ sont des lois NUVT.

Ainsi, dans l'algorithme MCMC principal, l'étape d'actualisation des $\boldsymbol{Z}_i$, $i = 1, \ldots, n$ qui initialement prenait la forme :

— Pour $i = 1, \ldots, n$, générer $\boldsymbol{Z}_i^{(t+1)}$ de la loi conditionnelle complète de $\boldsymbol{Z}_i$:

$$\boldsymbol{Z}_i^{(t+1)} \sim TN_p\big(\boldsymbol{X}_i\boldsymbol{\beta}^{(t)}, \sigma^2\big(\phi_i^{(t)}\big)^{-1}\boldsymbol{R}^{(t)}, \Omega_y\big).$$

est maintenant devenue :

— Pour $i = 1, \ldots, n$
  — Pour $j = 1, \ldots, p$, générer $Z_{ij}^{(t+1)}$ de la loi conditionnelle complète de Z_{ij} :

$$Z_{ij}^{(t+1)} \sim TN_1(\mu_{i,j.-j}^{(t)}, \sigma^{2(t)}_{i,j.-j}, \mathcal{R}_{ij})$$

où $\sigma^{2(t)}_{i,j.-j} = (H_{i,jj}^{(t)})^{-1}$, $\mu_{i,j.-j}^{(t)} = \mu_{ij}^{(t)} - (H_{i,jj}^{(t)})^{-1} \boldsymbol{H}_{i,j,-j}^{(t)} (\boldsymbol{z}_{i,-j}^{(t)} - \boldsymbol{\mu}_{i,-j}^{(t)})$,

$$H_{i,jj}^{(t)} = (\sigma^{-2}\phi_i^{(t)}(\boldsymbol{R}^{(t)})^{-1})_{jj}, \quad \boldsymbol{H}_{i,j,-j}^{(t)} = (\sigma^{-2}\phi_i^{(t)}(\boldsymbol{R}^{(t)})^{-1})_{j,-j},$$

$$\boldsymbol{\mu}_i^{(t)} = \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}, \quad \boldsymbol{z}_{i,-j}^{(t)} = (z_{i1}^{(t+1)}, \ldots, z_{i,j-1}^{(t+1)}, z_{i,j+1}^{(t)}, \ldots, z_{i,p}^{(t)})'$$

et la région de troncature $\mathcal{R}_{ij}$ étant l'intervalle $(0, +\infty)$ si $y_{ij} = 1$ et $(-\infty, 0)$ sinon.

CHAPITRE II

ANALYSE MULTIVARIÉE DES DONNÉES

2.1 Les mesures d'effet

Afin de résumer les résultats de nos analyses concernant l'association entre l'exposition et les réponses on suit l'approche de Hund *et al.* (2015), qui proposent deux types de mesures d'effet : le rapport de cotes (OR de l'anglais "odds ratio") et la différence de risque (ATE de l'anglais "average treatment effect"). L'estimation de ces deux mesures d'effet du traitement a été faite en adoptant le cadre contrefactuel de Neyman-Rubin (Rubin, 1974). Dans ce qui suit, on introduit la notation propre au cadre contrefactuel et on présente les expressions générales associées à chaque mesure dans un contexte de régression logistique multivariée bayésienne.

2.1.1 Notation

On commence par supposer l'existence de n unités expérimentales indexées par $i = 1, \ldots, n$, vues comme un échantillon aléatoire d'une population plus large. Pour chaque unité, on suppose aussi la connaissance du vecteur de réponses binaires $\boldsymbol{Y}_i = (Y_i^1, \ldots, Y_i^p)$ et la matrice de prédiction $\boldsymbol{X}_i = \text{diag}(\boldsymbol{X}_i^1, \ldots, \boldsymbol{X}_i^p)$ présentées à la fin de la section 1.2.3. On souligne que la notation présentée ici pour désigner les vecteurs de réponses, la matrice de prédiction et chaque sous-vecteur ligne est légèrement différente de celle de la section 1.2.3. Ce changement a été effectué pour

des raisons de clarté lors de la description mathématique des différentes variables modélisées dans cette étude. Dans chaque sous-vecteur ligne $\boldsymbol{X}_i^j$, $j = 1, \ldots, p$, de la matrice $\boldsymbol{X}_i$, on distingue la variable binaire associée à un traitement (ou exposition) E_i ($E_i = 0$ pour l'exposition de base et $E_i = 1$ pour l'exposition active) et l'ensemble des covariables $\boldsymbol{C}_i^j$. En outre, on définit $(\boldsymbol{Y}_i(0), \boldsymbol{Y}_i(1))$ comme la paire de réponses potentielles associées à la $i^{\text{ème}}$ unité expérimentale lorsque celle-ci reçoit l'exposition de base et l'exposition active, respectivement. Ainsi, le vecteur des réponses observées $\boldsymbol{Y}_i$ est relié à la paire des réponses potentielles $(\boldsymbol{Y}_i(0), \boldsymbol{Y}_i(1))$ par l'expression :

$$\boldsymbol{Y}_i \equiv \boldsymbol{Y}_i(E_i) = \begin{cases} \boldsymbol{Y}_i(0) & \text{si } E_i = 0, \\ \boldsymbol{Y}_i(1) & \text{si } E_i = 1. \end{cases}$$

Enfin, on précise que bien qu'on dérive les expressions de chaque mesure d'effet en supposant un traitement à deux niveaux dans cette section (variable binaire E_i), il est facile d'extrapoler celles-ci lorsque le traitement est à plusieurs niveaux (une variable binaire pour chaque modalité de traitement). Ceci est le cas dans notre étude, comme on le verra dans les sections ultérieures.

2.1.2 Les rapports de cotes

Dans la section 1.4, on a vu que le modèle t-link est une très bonne approximation au modèle de régression logistique multivarié original proposé par O'Brien et Dunson (2004). Ceci implique que, dans la pratique, les estimations des coefficients de régression $\boldsymbol{\beta}$ utilisant un modèle ou un autre sont presque identiques. Ainsi, étant donné que le modèle multivarié original a une structure marginale logistique, les coefficients de régression $\boldsymbol{\beta}$ estimés avec le modèle t-link correspondent approximativement aux logarithmes des différents rapports de cotes associationnels. C'est-à-dire :

$$\text{logit}(\text{P}(Y_i^j = 1|\boldsymbol{X}_i^j)) = \boldsymbol{X}_i^j\boldsymbol{\beta}^j = \beta_0^j + E_i\beta_E^j + \boldsymbol{C}_i^j\boldsymbol{\beta}_{\boldsymbol{C}}^j, \quad j = 1, \ldots, p \quad (2.1)$$

$$\text{et donc,} \quad OR^j = \frac{\text{P}(Y_i^j = 1|E_i = 1, \boldsymbol{C}_i^j)/(1 - \text{P}(Y_i^j = 1|E_i = 1, \boldsymbol{C}_i^j))}{\text{P}(Y_i^j = 1|E_i = 0, \boldsymbol{C}_i^j)/(1 - \text{P}(Y_i^j = 1|E_i = 0, \boldsymbol{C}_i^j))} = \exp(\beta_E^j).$$

On souligne que ces rapports de cotes sont interprétés conditionnellement aux covariables $\boldsymbol{C}$ et qu'ils sont supposés constants à travers tous les niveaux de ces variables.

Si certaines hypothèses sont respectées (Rubin, 1974), notamment que l'ensemble des variables confondantes est représenté dans $\boldsymbol{C}_i^j$, le rapport de cotes OR^j associationnel acquiert une interprétation causale et peut être exprimé en termes des réponses potentielles $Y_i^j(0)$ et $Y_i^j(1)$:

$$OR^j = \frac{\text{E}[Y_i^j(1)|\boldsymbol{C}_i^j]/(1 - \text{E}[Y_i^j(1)|\boldsymbol{C}_i^j])}{\text{E}[Y_i^j(0)|\boldsymbol{C}_i^j]/(1 - \text{E}[Y_i^j(0)|\boldsymbol{C}_i^j])} = \exp(\beta_E^j).$$

En suivant une approche bayésienne, l'estimation de l'effet de l'exposition E sur chaque réponse binaire Y^j au moyen du rapport de cotes OR^j, $j = 1, \ldots, p$, est basée sur la loi *a posteriori* de OR^j. Dans la pratique, on obtient une estimation de cette loi *a posteriori* au moyen d'un échantillon aléatoire $\{(\beta_E^j)^{(1)}, (\beta_E^j)^{(2)}, \ldots, (\beta_E^j)^{(T)}\}$ généré de la loi *a posteriori* du coefficient β_E^j sous le modèle t-link. La moyenne et/ou la médiane empiriques ainsi que des intervalles de crédibilité peuvent être obtenus afin de résumer la loi de OR^j.

2.1.3 Les différences de risque : ATE

Un avantage de l'approche bayésienne sur les méthodes d'estimation classiques est qu'elle facilite l'inférence pour les fonctions des paramètres du modèle, ce qui est le cas pour les différences de risque individuel ou spécifique à chaque maladie et les différences de risque multiple.

On définit la différence de risque spécifique à une maladie (ATE^j, $j = 1, \ldots, p$) comme la différence de risque de subir la maladie si toute la population était traitée versus non traitée. Exprimée en termes contrefactuels, cette mesure de l'effet d'un traitement prend la forme $ATE^j = \mathrm{E}[Y^j(1) - Y^j(0)] = \mathrm{E}_{\boldsymbol{C}}[\mathrm{E}[Y^j(1) - Y^j(0)|\boldsymbol{C}]] = \mathrm{E}_{\boldsymbol{C}}[(\mathrm{P}(Y(1) = 1) - \mathrm{P}(Y(0) = 1))|\boldsymbol{C}]$. Sous l'hypothèse que l'échantillon est représentatif de la population d'intérêt, on peut définir un ATE^j au moyen de l'estimation d'un ATE^j empirique conditionnel à la distribution empirique de l'ensemble de covariables associées aux participants, $ATE^j = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i^j(1) - Y_i^j(0)|\boldsymbol{C}_i^j] = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{P}(Y_i^j(1) = 1|\boldsymbol{C}_i^j) - \mathrm{P}(Y_i^j(0) = 1|\boldsymbol{C}_i^j))$. Si certaines conditions concernant l'identifiabilité d'effets causaux sont respectées, il est possible de montrer que $\mathrm{P}(Y_i^j(e) = 1|\boldsymbol{C}_i^j) = \mathrm{P}(Y_i^j = 1|E_i = e, \boldsymbol{C}_i^j)$, $e = 0, 1$. En combinant ce résultat avec l'expression (2.1) on obtient :

$$\mathrm{P}(Y_i^j(e) = 1|\boldsymbol{C}_i^j) = \frac{\exp(\beta_0^j + e\beta_E^j + \boldsymbol{C}_i^j\boldsymbol{\beta}_{\boldsymbol{C}}^j)}{1 + \exp(\beta_0^j + e\beta_E^j + \boldsymbol{C}_i^j\boldsymbol{\beta}_{\boldsymbol{C}}^j)}, \; e = 0, 1. \tag{2.2}$$

L'estimation bayésienne des ATE^j, $j = 1, \ldots, p$, est basée sur les lois *a posteriori* des différences de risque potentielles $\mathrm{P}(Y_i^j(e) = 1|\boldsymbol{C}_i^j)$, $e = 0, 1$, associées aux participants et leur marginalisation ultérieure. Plus précisément, on échantillonne de la loi *a posteriori* des coefficients de la régression $\boldsymbol{\beta}$ sous le modèle t-link et, pour chaque participant, on génère un échantillon des différences de risque contrefactuelles au moyen de l'expression (2.2). Finalement, on caractérise la loi *a posteriori* associée au ATE^j en calculant les différences de risque obtenues pour chaque élément de l'échantillon de $\boldsymbol{\beta}$.

On souligne que les estimations de ATE^j empirique ne sont pas de bonnes estimations de ATE^j pour des populations qui n'ont pas des caractéristiques similaires à l'échantillon de travail. Néanmoins, comprendre comment les risques de subir une maladie ou un trouble de santé changent en fonction du traitement reçu dans

l'échantillon de travail reste d'intérêt.

Le modèle de régression multivarié t-link, à l'opposé d'un ensemble de régressions logistiques standards, permet aussi d'estimer la probabilité de développer conjointement une combinaison quelconque de troubles médicaux selon le niveau de traitement. Ainsi, on peut estimer les différences de risque associées au traitement pour chaque combinaison des troubles cardiométaboliques. Formellement, les différences de risque multiple marginalisées sur l'ensemble des covariables de l'échantillon sont définies comme :

$$ATE^{\boldsymbol{l}} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{P}(\boldsymbol{Y}_i(1) = \boldsymbol{l}|\boldsymbol{C}_i) - \mathrm{P}(\boldsymbol{Y}_i(0) = \boldsymbol{l}|\boldsymbol{C}_i))$$

où $\boldsymbol{Y}_i(e) = (Y_i^1(e), Y_i^2(e), \ldots, Y_i^p(e))$, $\boldsymbol{l} = (l_1, l_2, \ldots, l_p)$ et $l_j, e \in \{0,1\}$, $j = 1, 2, \ldots, p$.

Sous le modèle t-link, les probabilités potentielles sont calculées en utilisant l'expression :

$$\mathrm{P}(\boldsymbol{Y}_i(e) = \boldsymbol{l}|\boldsymbol{C}_i) = \int_{\Omega_{\boldsymbol{l}}} \mathcal{T}_p(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}, \nu)\,\mathrm{d}\boldsymbol{z}_i, \tag{2.3}$$

où $\boldsymbol{X}_i = \mathrm{diag}(\boldsymbol{X}_i^1, \ldots, \boldsymbol{X}_i^p)$, $\boldsymbol{X}_i^j = (1, e, \boldsymbol{C}_i^j)$ et $\Omega_{\boldsymbol{l}} = \{\boldsymbol{z}_i \in \mathbb{R}^4 : z_{ij} > 0 \text{ si } l_j = 1, z_{ij} \leq 0 \text{ sinon }, j = 1, \ldots, p\}$.

L'estimation bayésienne des différences de risque multiple ($ATE^{\boldsymbol{l}}$) suit une démarche analogue à celle présentée pour l'estimation des différences individuelles (ATE^j) : on échantillonne de la loi *a posteriori* des paramètres du modèle ($\boldsymbol{\beta}$ et $\boldsymbol{R}$) et, pour chaque participant, on induit un échantillon des différences de risque multiple au moyen de l'expression (2.3). Enfin, on marginalise sur l'ensemble des participants pour obtenir des réalisations de la loi *a posteriori* de $ATE^{\boldsymbol{l}}$.

Le calcul des différences de risque pour chaque combinaison des troubles cardiométaboliques permet aussi d'estimer la différence en risque de subir un nombre fixé k

de troubles ou plus si l'échantillon entier était traité versus non traité ($ATE^{N\geq k}$). On note par $Y_i^{N\geq k}(e)$, $k = 1, \ldots, p$, la réponse binaire potentielle prenant la valeur 1 lorsque, en fixant le traitement au niveau e, le $i^{\text{ème}}$ participant subit k troubles cardiométaboliques ou plus et 0 sinon. Alors, la différence de risque $ATE^{N\geq k}$ prend la forme :

$$ATE^{N\geq k} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{P}(Y_i^{N\geq k}(1) = 1|\boldsymbol{C}_i) - \mathrm{P}(Y_i^{N\geq k}(0) = 1|\boldsymbol{C}_i)).$$

Comme dans le cas des autres différences de risque, l'estimation des $ATE^{N\geq k}$ se base sur sa loi *a posteriori*, qui est facilement approximée si les différences ATE^l ont été préalablement estimées.

2.2 Description des données

2.2.1 La cohorte PÉTALE

Les survivants de la leucémie aiguë lymphoblastique de l'enfant (LALe) qui ont fait l'objet de notre étude ont été recrutés comme participants dans l'étude PÉTALE menée par le Centre hospitalier universitaire Sainte-Justine (CHUSJ) à Montréal. L'étude PÉTALE est un projet de recherche multidisciplinaire visant trois objectifs. Le premier objectif est de caractériser les effets indésirables à long terme du traitement (EIL) les plus communs pour une population de jeunes survivants de la LALe, concrètement : syndrome métabolique, cardiotoxicité, complications osseuses et problèmes concernant la qualité de vie. Le deuxième objectif est d'identifier des possibles biomarqueurs cliniques, biochimiques et génétiques pouvant aider à prédire l'apparition des EIL. Finalement, le troisième objectif est de proposer des recommandations et des interventions afin d'atténuer le risque d'apparition et la sévérité des EIL.

La cohorte PÉTALE est composée de jeunes survivants de LALe qui ont été traités

au CHUSJ et qui respectent les critères d'éligibilité suivants :

- Être de descendence européenne et âgé de moins de 19 ans au moment du diagnostic de LALe entre 1987 et 2010.
- Avoir été traité au CHUSJ selon les protocoles associés au traitement de LALe du Dana-Farber Cancer Institute (DFCI-ALL) : 87-01, 91-01, 95-01, 2000-01, 2005-01.
- Avec 5 ans ou plus de temps écoulé depuis le diagnostic.
- Avoir donné son consentement éclairé si $\geq$ 18 ans, ou avoir donné son assentiment et les consentements éclairés de ses parents si $<$ 18 ans.
- Ne pas avoir souffert d'une LALe récidivante ou réfractaire, ni subi une greffe de cellules souches, ni reçu des médicaments ostéotoxiques, et n'avoir aucune maladie osseuse congénitale.

Entre 2012 et 2016, les participants de l'étude PÉTALE ont été soumis à des tests et analyses cliniques, biochimiques et psychosociaux exhaustifs et spécifiques à chaque catégorie de EIL afin de bien caractériser l'ensemble de ceux-ci. Une description détaillée des méthodes et équipements utilisés est présentée dans Marcoux, *et al.*, (2016).

Dans notre recherche visant à estimer l'impact du traitement de la LALe sur la probabilité de développer les facteurs de risque cardiométabolique (obésité, résistance à l'insuline, (pré)-hypertension et dyslipidémie), on a travaillé avec un sous-ensemble de survivants de la cohorte PÉTALE pour lesquels on dispose des données complètes concernant le traitement, les variables de confusion potentielles et les facteurs de risque cardiométabolique.

2.2.2 Les variables

Les variables réponses

Les quatre complications cardiométaboliques : obésité, résistance à l'insuline, (pré)-hypertension et dyslipidémie constituent l'ensemble des variables réponses binaires (oui/non) dans notre étude; chacune de ces variables a été obtenue à partir de différentes mesures associées à la santé cardiométabolique de la façon suivante.

L'obésité a été établie chez un individu lorsqu'au moins un de ces deux facteurs était présent à l'entrevue : obèse selon l'indice de masse corporelle IMC= poids/taille^2 (kg/m^2) ($\geq$ 30 kg/m^2 pour adultes, $\geq$ 97^e centile pour enfants) ou tour de taille élevé ($\geq$ 102 cm pour hommes, $\geq$ 88 cm pour femmes et $\geq$ 95^e centile pour enfants) (Katzmarzyk, 2004; Van den Broeck *et al.*, 2009).

La résistance à l'insuline a été définie à partir de trois facteurs : glycémie à jeun élevée ($\geq$ 6.1 mmol/L), pourcentage d'hémoglobine glyquée (facteur HbA1c) élevée ($>$ 6%) ou valeurs estimées élevées selon le modèle d'homéostasie de la résistance à l'insuline (HOMA-IR)[insuline (mIU/L) $\times$ glucose (mmol/L)/22.5]($\geq$ 2.86 pour adultes, $\geq$ 95^e centile pour enfants) (Chen *et al.*, 2004; Allard *et al.*, 2003).

La pression artérielle a été mesurée le matin au bras droit des participants assis et au repos. La (pré)-hypertension a été établie en suivant les recommandations actuelles pour adultes (Paradis *et al.*, 2010; Stern, 2013) : on considère qu'un individu souffre de (pré)-hypertension lorsque sa pression artérielle systolique est $\geq$ 130 mmHg ou que sa pression artérielle diastolique est $\geq$ 85 mmHg ($\geq$ 90^e centile selon âge et poids pour enfants). Si l'individu est sous antihypertenseurs il est aussi considéré hypertendu.

La dyslipidémie a été évaluée en sérum à jeun. Un individu a été classé comme

souffrant de dyslipidémie s'il avait au moins un de ces quatre facteurs à l'entrevue : taux de cholestérol LDL élevé ($\geq$ 3.4 mmol/L pour adultes, $\geq$ 3.36 mmol/L pour enfants), taux de triglycérides élevé ($\geq$ 1.7 mmol/L pour adultes, $\geq$ 1.47 mmol/L pour enfants), taux de cholestérol HDL bas ($<$ 1.03 mmol/L pour hommes et enfants, $<$ 1.3 mmol/L pour femmes) ou suivant un traitement pour faire baisser le taux de lipides et/ou cholestérol (Genest *et al.*, 2009).

Enfin, on souligne que, au vu de la complexité dans les définitions des variables réponses, il est naturel de définir celles-ci comme des variables binaires plutôt que continues.

La variable d'exposition

La variable d'exposition a été définie à partir des doses cumulatives reçues de CS et de RT. Les doses cumulatives de CS ont été calculées comme la somme, en doses équivalentes de prednisone, des doses cumulées normalisées par la surface corporelle (mg/m^2) de dexamethasone, methylprednisolone et prednisone reçues pendant les trois phases du traitement : induction, consolidation et entretien. On a utilisé la médiane de la distribution empirique des doses cumulées (8494 mg/m^2) pour catégoriser l'ensemble des survivants selon s'ils ont reçus une dose cumulée faible de CS (DF) ou une dose cumulée élevée de CS (DE). En ce qui concerne l'exposition à la radiothérapie, étant donné que la plupart des survivants ont reçu, soit une exposition de 18Gy, soit aucune exposition, on l'a quantifiée au moyen d'une variable binaire (oui/non). Ainsi, la variable exposition associée au traitement est une variable catégorielle à trois niveaux : 1) DF/non RT, 2) DF/RT, 3) DE/RT, quantifiant ainsi l'intensité du traitement combiné de CS et RT. La catégorie d'exposition DE/non RT a été exclue puisque seul un nombre très petit de survivants ont reçu une dose élevée de CS sans être exposés à la RT.

Les variables d'ajustement

On s'intéresse à l'impact de l'exposition sur le risque de développer un ensemble de complications cardiométaboliques quelconque. Ainsi, l'ajustement par des variables potentiellement confondantes est essentielle si on veut obtenir des estimations de l'effet de l'exposition qui puissent être pertinentes dans une optique d'intervention médicale. Les variables d'ajustement considérées dans nos analyses ont été le sexe, l'âge au diagnostic (années), le temps depuis le diagnostic (années) et la concentration de cellules blanches au diagnostic ($\times 10^9$/L), un proxy de la sévérité de la maladie noté par WBC. Les variables d'ajustement sélectionnées sont toutes des variables potentiellement confondantes puisque chacune est associée au niveau d'exposition et/ou aux réponses cardiométaboliques étudiées. En particulier, le sexe, l'âge et le WBC ont été utilisées pour classifier les patients dans les groupes de risque de rechute standard et élevé dans les protocoles DFCI-CALL 1987-01 et 2005-01.

2.3 Analyses

L'analyse bayésienne de l'effet de l'exposition sur la prévalence des complications cardiométaboliques repose sur le modèle *t-link* d'O'Brien et Dunson (2004) présenté à la section 1.4. L'implantation de l'algorithme MCMC décrit à la section 1.5 a permis la simulation efficace des lois *a posteriori* des paramètres $\boldsymbol{\beta}$ et $\boldsymbol{R}$ et l'obtention des estimations d'effets en ajustant pour différentes covariables.

Dans ce qui suit, on décrit précisément la modélisation des variables d'intérêt au moyen du modèle *t-link*, l'implantation de l'algorithme MCMC ainsi que les différentes simulations et analyses effectuées.

2.3.1 Modélisation des variables

Soit $\boldsymbol{Y}_i = (Y_i^O, Y_i^I, Y_i^H, Y_i^D)$ le vecteur de réponses binaires (1/0) pour le $i^{\text{ème}}$ survivant où O, I, H et D désignent l'obésité, la résistance à l'insuline, la (pré)-hypertension et la dyslipidémie, respectivement. Soient aussi $T_{1,i}$ et $T_{2,i}$ des variables indicatrices codant les niveaux de traitement 2^e et 3^e, respectivement, versus le niveau de base (T_0 : DF/non RT) pour ce survivant et $\boldsymbol{C}_i$ un vecteur ligne de variables d'ajustement (sexe, âge au diagnostic, temps depuis le diagnostic et concentration de cellules blanches (WBC) au diagnostic). Alors, sous le modèle *t-link*, ces variables sont reliées par les expressions :

$$Y_i^j = \mathbb{1}(Z_i^j > 0) \quad j \in \{O, I, H, D\},$$
$$\boldsymbol{Z}_i|\boldsymbol{\beta}, \boldsymbol{R} \sim \mathcal{T}_4(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}, \nu),$$

où $\boldsymbol{X}_i\boldsymbol{\beta} = \text{diag}(\boldsymbol{X}_i^O, \boldsymbol{X}_i^I, \boldsymbol{X}_i^H, \boldsymbol{X}_i^D)((\boldsymbol{\beta}^O)', (\boldsymbol{\beta}^I)', (\boldsymbol{\beta}^H)', (\boldsymbol{\beta}^D)')'$ et
$\boldsymbol{X}_i^j\boldsymbol{\beta}^j = \beta_0^j + T_{1,i}\beta_1^j + T_{2,i}\beta_2^j + \boldsymbol{C}_i\boldsymbol{\beta}_{\boldsymbol{C}}^j$.

On rappelle que, afin de bien préserver la structure marginale logistique du modèle original d'O'Brien et Dunson (2004), on a fixé le paramètre ν à 7.3 et σ^2 à $\pi^2(\nu - 2)/3\nu$. On souligne aussi que, bien que la notation choisie pour les sous-vecteurs de la matrice $\boldsymbol{X}_i$ permet de différencier ceux-ci, dans notre étude on a $\boldsymbol{X}_i^O = \boldsymbol{X}_i^I = \boldsymbol{X}_i^H = \boldsymbol{X}_i^D$ puisqu'on a utilisé le même ensemble de variables d'ajustement pour chaque complication cardiométabolique.

2.3.2 L'algorithme MCMC et les simulations

Afin d'estimer la loi *a posteriori* des paramètres $\boldsymbol{\beta}$ et $\boldsymbol{R}$, on a implanté l'algorithme MCMC présenté à la section 1.5. Cet algorithme présuppose l'indépendance *a priori* des paramètres, c'est-à-dire $p(\boldsymbol{\beta}, \boldsymbol{R}) = p(\boldsymbol{\beta})p(\boldsymbol{R})$, ainsi que la normalité des coefficients de régression : $\boldsymbol{\beta} \sim \mathcal{N}_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. Plus précisément, on a posé $p(\boldsymbol{\beta}) = \prod_j \mathcal{N}(\boldsymbol{\beta}^j|\boldsymbol{\beta}_0^j, \boldsymbol{\Sigma}_0)$, où $\boldsymbol{\beta}_0^j = \mathbf{0}$ et $\boldsymbol{\Sigma}_0 = \text{diag}(1000, 4, \ldots, 4)$, pour $j \in$

$\{O, I, H, D\}$, et $p(\boldsymbol{R})$ comme étant uniforme sur l'espace des matrices de corrélation. Un prototype de l'algorithme MCMC a été développé en langage R à l'aide de l'environnement de développement intégré RStudio. Ultérieurement, dans le but d'augmenter la performance de l'algorithme, le programme a été réécrit en C++ au moyen des paquets informatiques Rcpp et RcppArmadillo. Le code informatique associé à l'algorithme MCMC ainsi que le code pour la lecture et la manipulation des données se trouvent à l'annexe C. Également, dans l'annexe A on trouve les résultats concernant la validation informatique de l'algorithme au moyen de deux jeux de données simulés.

Un élément critique dans l'implantation de l'algorithme a été l'ajustement de la loi de proposition de l'étape Metropolis-Hastings, loi qui permet d'actualiser les six coefficients uniques associés à la matrice de corrélation $\boldsymbol{R}$ (voir section 1.4, page 28). En effet, tel que commenté à la section 1.3, le choix du paramètre d'échelle d'une loi de proposition a un impact important sur le comportement de la chaîne, en particulier sur la dépendance entre réalisations et par conséquence, sur la vitesse de convergence et la capacité de bien explorer l'espace d'états. Dans notre algorithme MCMC, le paramètre d'échelle à ajuster était la matrice de covariances $\boldsymbol{\Omega}$ (6×6) d'une loi de proposition normale multivariée. À cause de la dimension du problème, l'ajustement complet de la matrice des covariances $\boldsymbol{\Omega}$ par simple essai-erreur n'était pas viable. Or, certains résultats suggèrent qu'un choix optimal consiste à prendre comme matrice de covariances celle de la loi cible multipliée par un facteur d'échelle particulier (Rosenthal, 2010). Ainsi, inspirés par ceux-ci, on a décidé de prendre comme choix pour $\boldsymbol{\Omega}$ la matrice de covariances empirique d'un échantillon s'approchant de la loi *a posteriori* de $\boldsymbol{R}$ (loi cible) multipliée par un facteur $(= 0.08)$ ajusté par essai-erreur. Cet échantillon a été obtenu à partir d'une simulation préliminaire assez longue pour extraire un nombre suffisamment grand de réalisations faiblement corrélées de la loi *a posteriori* de $\boldsymbol{R}$.

On a remarqué que la valeur d'échelle avec laquelle on a obtenu les meilleurs résultats (= 0.08) est assez différente de celle proposée dans Rosenthal (2010) pour une loi cible normale multivariée de dimension 6 ($\simeq$ 1). Un facteur qui pourrait expliquer cet écart est le fait que le support de notre loi cible est une région relativement petite de l'hypercube $[-1,1]^6$ associé à l'espace des matrices symétriques de dimension 4×4 générées avec notre loi de proposition (Rousseeuw et Molenberghs, 1994).

Afin de vérifier la plausible stabilité et convergence des chaînes simulées avec l'algorithme MCMC, on a réalisé un diagnostic de convergence en comparant les distributions des réalisations de trois chaînes selon trois conditions initiales différentes. Les autocorrélations entre réalisations pour chaque paramètre du modèle ont été aussi évaluées. Les détails sur le diagnostic de convergence ainsi que les résultats correspondants obtenus sont présentés à l'annexe B.

Les analyses concernant le lien exposition-réponses cardiométaboliques ont été effectuées avec trois modèles : modèle sans variables d'ajustement (analyse brute), modèle ajusté et modèle ajusté sans WBC. Pour chaque modèle, on a généré 1050000 réalisations de la loi *a posteriori* de $(\boldsymbol{\beta}, \boldsymbol{R})$. Bien que le processus d'ajustement de la loi de proposition décrit antérieurement a permis d'améliorer la convergence pour $\boldsymbol{R}$, celle-ci restait encore lente et l'autocorrélation entre réalisations consécutives trop élevée. Par conséquent, on a écarté les 50000 premières valeurs ('burn-in') afin de réduire l'influence de celles-ci sur les statistiques obtenues (moyenne, médianes, IC). De plus, l'autocorrélation a été considérablement atténuée en gardant seulement toutes les 100 itérations ('thinning' = 100). Ainsi, pour chaque modèle, le nombre total d'itérations disponibles pour faire de l'inférence était de 10000.

L'effet de chaque niveau de traitement (T_l vs $T_0, l = 1, 2$) sur l'ensemble des ré-

ponses cardiométaboliques a été estimé à l'aide des mesures d'effet présentées à la section 2.1 : les rapports de cotes (OR^j, $j \in \{O, I, H, D\}$), les différences de risque individuel ou spécifique à chaque maladie (ATE^j, $j \in \{O, I, H, D\}$) et les différences en risque de subir un nombre fixé k de troubles ou plus ($ATE^{N \geq k}$, $k = 1, \ldots, 4$). Pour chaque modèle, des estimations des lois *a posteriori* des mesures d'effet ont étés obtenues au moyen d'échantillons de taille 10000 de $(\boldsymbol{\beta}, \boldsymbol{R})$ en suivant les démarches décrites à la section 2.1. La moyenne ainsi que l'intervalle de crédibilité à 95% ont été utilisés afin de résumer les lois. De plus, à cause de l'asymétrie des lois *a posteriori* des ORs, la médiane a aussi été présentée pour cette mesure d'association. Enfin, puisque nous pouvons interpréter les coefficients du modèle multivarié de façon marginale pour chacune des réponses, on a cru intéressant de comparer les estimations des OR obtenus de ce modèle à ceux de régressions logistiques standards appliquées à chaque complication cardiométabolique séparément.

On souligne que l'obtention des $ATE^{N \geq k}$, $k = 1, \ldots, 4$, pour chaque niveau d'exposition demande le calcul des différences multiples ($ATE^{\boldsymbol{l}}$) au moyen des intégrales (2.3). L'évaluation de ces intégrales, gourmandes en temps de calcul, a été réalisée grâce à la fonction *pmvt* du logiciel R. Plus précisément, la fonction *pmvt* nous a permis d'évaluer la probabilité que le vecteur auxilaire $\boldsymbol{Z}_i$ distribué selon une loi de Student multivariée tombe dans l'hyper-rectangle $\mathcal{R}_i = \prod_{j=1}^{4} \mathcal{R}_i^j$ où $\mathcal{R}_i^j = (0, +\infty)$ ou $\mathcal{R}_i^j = (-\infty, 0)$, selon les valeurs prises par les variables réponses binaires Y_i^j.

Par la suite, des analyses supplémentaires ciblant l'association entre les variables d'ajustement et les réponses cardiométaboliques pour les groupe de survivants ayant reçu le niveau de traitement élevé (DE/RT) ont été effectuées. Finalement, des analyses de sensibilité pour l'association entre WBC et la résistance à l'insuline

ont été faites au moyen de modèles de régression logistique standards où la variable WBC a été dichotomisée selon une suite de valeurs de seuil définies selon les valeurs observées de la variable dans notre cohorte.

Dans la section suivante, on présente et interprète les résultats obtenus en suivant la démarche expérimentale que nous venons de décrire.

2.4 Résultats

Des statistiques descriptives de la cohorte PÉTALE sont présentées à la table 2.1. On souligne que les survivants de notre cohorte sont en majorité des femmes (100/180 ou 55.6%). L'âge moyen au diagnostic de la LALe était de 5.8 ans (ET (écart-type) = 4.0), tandis que le temps moyen écoulé entre le diagnostic et l'entrevue est de 15.2 ans (ET = 4.8). Parmi les 180 survivants, 71 (39.4%) ont reçu de faibles doses de corticostéroïdes sans radiothérapie (DF/non RT), 24 (13.3%) de faibles doses avec radiothérapie (DF/RT) et 85 (47.2%) des doses élevées avec radiothérapie (DE/RT). Le taux de prévalence des quatre complications cardiométaboliques à l'entrevue était assez élevé, allant de 10% pour la (pré)-hypertension à 38.9% pour la dyslipidémie.

On souligne aussi des déséquilibres marqués concernant la distribution des covariables Genre et WBC entre les différents niveaux de traitement. Par ailleurs, on remarque que les survivants ayant reçu le traitement de base (DF/non RT) sont les plus jeunes, en plus de présenter les plus basses valeurs de WBC et les plus courts temps écoulés depuis le diagnostic.

2.4.1 Résultats modèle multivarié

Des résultats bruts (non ajustés) et ajustés concernant l'effet de chaque niveau de traitement sont présentés aux tables 2.2 et 2.3 pour chaque complication cardio-

Tableau 2.1 Caractéristiques marginales et stratifiées (par niveau de traitement) de la cohorte de survivants de LALe.

	Cohorte complète ($n = 180$)	DF/non RT ($n = 71$)	DF/RT ($n = 24$)	DE/RT ($n = 85$)
Obésité	55 (30.6%)	19 (26.8%)	8 (33.3%)	28 (32.9%)
Résistance à l'insuline	30 (16.7%)	8 (11.3%)	5 (20.8%)	17 (20.0%)
(Pré-)hypertension	18 (10.0%)	4 (5.63%)	2 (8.33%)	12 (14.1%)
Dyslipidémie	70 (38.9%)	17 (23.9%)	12 (50.0%)	41 (48.2%)
Genre (mâle)	80 (44.4%)	23 (32.4%)	18 (75%)	39 (45.8%)
Âge au diagnostic (années)	5.83 (3.97)	4.66 (2.20)	6.07 (4.95)	6.74 (4.56)
Temps depuis diagnostic (années)	15.2 (4.78)	14.21 (4.39)	16.5 (4.41)	15.7 (5.07)
WBC ($\times 10^9$/L)	29.8 (48.8)	8.30 (7.29)	23.9 (35.9)	49.5 (62.4)

Les résultats sont présentés en % (entre parenthèses) pour les variables binaires et avec la moyenne et l'écart-type (entre parenthèses) pour les variables continues. DF : dose faible de corticostéroïdes; DE : dose élevée de corticostéroïdes; RT : radiothérapie; WBC : concentration de cellules blanches.

métabolique individuellement. On souligne que, sous les trois modèles multivariés, le niveau de traitement DE/RT (versus le niveau de base DF/non RT) est significativement associé à une augmentation de la prévalence de la dyslipidémie chez les survivants, et cela selon les deux mesures d'effet utilisées : le rapport de cotes (OR^D) et la différence de risque individuel (ATE^D). En particulier, dans le modèle ajusté incluant le WBC, les médianes et intervalles de crédibilité pour ces mesures sont $OR^D = 2.49$ (IC à 95% : $1.16, 5.48$) et $ATE^D = 0.19$ (IC à 95% : $0.03, 0.35$), respectivement. Aucune autre association entre un niveau de traitement et une complication cardiométabolique n'a pu être considérée comme statistiquement significative (tous les intervalles de crédibilité des ORs et des ATEs restants incluent leur valeur de référence, soit 1 ou 0, respectivement). On remarque que l'inclusion de la variable WBC dans le modèle a une faible influence sur l'estimation de l'effet du traitement sur la (pré)-hypertension et la dyslipidémie individuellement. En

effet, les valeurs des moyennes et médianes ainsi que les intervalles de crédibilité estimés pour OR^H, OR^D, ATE^H et ATE^D n'exhibent pas de changements substantiels lorsqu'on passe du modèle ajusté sans inclure WBC au modèle ajusté incluant WBC (modèle ajusté complet). Par contre, on observe une situation différente pour l'obésité et la résistance à l'insuline : lorsqu'on inclut WBC dans le modèle, les valeurs associées aux mesures d'effet OR^O, OR^I, ATE^O et ATE^I sont sensiblement atténuées. Ces résultats suggèrent que WBC est une variable de confusion pour ces deux réponses cardiométaboliques. En fait, WBC est un très fort déterminant du niveau de traitement reçu et l'association entre WBC et la résistance à l'insuline a été trouvée statistiquement significative avec un $OR = 1.17$ pour un incrément de 10 unités de WBC (IC à 95% : $1.07, 1.29$).

En examinant les éléments estimés de la matrice $\boldsymbol{R}$ dans le modèle ajusté complet, on remarque une corrélation résiduelle positive entre l'obésité et la dyslipidémie ($\rho = 0.30$, IC à 95% : $0.08, 0.51$) et une autre encore plus marquée entre l'obésité et la résistance à l'insuline ($\rho = 0.68$, IC à 95% : $0.48, 0.84$). Ces deux résultats viennent souligner ce que d'autres études ont mis en évidence, à savoir l'existence de liens causaux entre l'obésité et la résistance à l'insuline (Hardy *et al.*, 2012; Kahn et Flier, 2000), ainsi qu'entre l'obésité et la dyslipidémie (Klop *et al.*, 2013).

La table 2.4 présente des résultats concernant l'impact global du traitement sur l'ensemble des complications cardiométaboliques. En particulier, on a estimé les différences de risque cumulé $ATE^{N \geq k}$, $k = 1, \ldots, 4$, pour chaque niveau de traitement sur la base des deux modèles ajustés. On voit que, dans le modèle complètement ajusté, le niveau de traitement le plus élevé (DE/RT) est significativement associé à une augmentation de 0.15 du risque de développer au moins une complication métabolique ($N \geq 1$) en comparaison au niveau de traitement de base (DF/non RT). La valeur de cette différence de risque est expliquée principalement par les différences de risque multiple (ATE^l) de deux combinaisons spéci-

Tableau 2.2 Rapports de cotes (ORs) bruts et ajustés associés au traitement pour les facteurs de risque cardiométabolique individuels. *Les estimations ponctuelles sont la moyenne/médiane a posteriori obtenues avec* 10000 *réalisations.*

	$OR^{\dagger}$ (Intervalle de crédibilité 95%)††		
	Brut	Ajusté incluant WBC	Ajusté excluant WBC
Obésité			
DF/RT	1.41/1.25 (0.44, 3.40)	1.75/1.50 (0.47, 4.37)	1.85/1.59 (0.53, 4.72)
DE/RT	1.41/1.33 (0.68, 2.61)	1.18/1.09 (0.49, 2.40)	1.46/1.36 (0.65, 2.85)
Résistance à l'insuline			
DF/RT	2.01/1.68 (0.49, 5.41)	1.71/1.39 (0.34, 5.03)	2.04/1.66 (0.43, 6.03)
DE/RT	2.12/1.91 (0.82, 4.60)	0.98/0.85 (0.29, 2.46)	2.09/1.85 (0.75, 4.84)
(Pré)-hypertension			
DF/RT	1.80/1.32 (0.21, 6.30)	1.29/0.93 (0.15, 4.69)	1.25/0.89 (0.14, 4.51)
DE/RT	3.07/2.55 (0.93, 8.44)	3.03/2.40 (0.71, 8.85)	3.03/2.42 (0.74, 9.07)
Dyslipidémie			
DF/RT	3.34/2.97 (1.17, 7.64)	2.70/2.36 (0.84, 6.65)	2.63/2.30 (0.83, 6.40)
DE/RT	2.99/2.83 (1.45, 5.53)	2.70/2.49 (1.16, 5.48)	2.51/2.34 (1.15, 4.81)

† : Le traitement de base est DF/non RT.
†† : Les bornes des intervalles sont les percentiles empiriques 2.5 % et 97.5 %.
DF : dose faible de corticostéroïdes; DE : dose élevée de corticostéroïdes; RT : radiothérapie; WBC : concentration de cellules blanches.

fiques de troubles métaboliques : $(Y^O = 0, Y^I = 0, Y^H = 0, Y^D = 1)$ et $(Y^O = 0, Y^I = 0, Y^H = 1, Y^D = 1)$. Pour la première combinaison, la valeur estimée de l'$ATE^{\boldsymbol{l}}$, $\boldsymbol{l} = (0, 0, 0, 1)$, est 0.10 (IC à 95% : $-0.00, 0.21$) tandis que pour la deuxième l'$ATE^{\boldsymbol{l}}$, $\boldsymbol{l} = (0, 0, 1, 1)$, on a obtenu 0.02 (IC à 95% : $0.00, 0.06$). On souligne que, pour cette dernière combinaison, l'estimation de l'$ATE^{\boldsymbol{l}}$ est statistiquement significative, ce qui indique que le niveau de traitement le plus élevé (vs traitement de base) semble augmenter légèrement (0.02) le risque de présenter une dyslipidémie et une (pré)-hypertension *simultanément* sans obésité ni résistance à

Tableau 2.3 Différences de risque individuel (ATEs) brutes et ajustées associées au traitement pour les facteurs de risque cardiométabolique. *Les estimations ponctuelles sont les moyennes a posteriori obtenues avec* 10000 *réalisations.*

	ATE (Intervalle de crédibilité 95%)		
	Brut	Ajusté incluant WBC	Ajusté excluant WBC
Obésité			
DF/RT	0.05 (-0.15, 0.27)	0.08 (-0.14, 0.31)	0.09 (-0.12, 0.32)
DE/RT	0.06 (-0.08, 0.19)	0.01 (-0.14, 0.17)	0.06 (-0.08, 0.20)
Résistance à l'insuline			
DF/RT	0.07 (-0.07, 0.24)	0.04 (-0.11, 0.22)	0.06 (-0.08, 0.25)
DE/RT	0.08 (-0.03, 0.19)	-0.02 (-0.14, 0.09)	0.07 (-0.04, 0.18)
(Pré)-hypertension			
DF/RT	0.02 (-0.07, 0.16)	-0.00 (-0.09, 0.11)	-0.00 (-0.09, 0.10)
DE/RT	0.08 (-0.01, 0.17)	0.07 (-0.03, 0.17)	0.07 (-0.03, 0.16)
Dyslipidémie			
DF/RT	0.24 (0.03, 0.45)	0.18 (-0.04, 0.41)	0.18 (-0.04, 0.40)
DE/RT	0.23 (0.08, 0.37)	0.19 (0.03, 0.35)	0.18 (0.03, 0.33)

voir table 2.2 pour les notes et légendes.

l'insuline. Par ailleurs, on note que toutes les $ATE^{N\geq k}$, $k = 1, \ldots, 4$, associées au niveau de traitement DE/RT sont positives et que, dans le modèle ajusté excluant WBC, celles-ci sont significativement différentes de zéro. Ceci pourrait indiquer un potentiel effet global nocif du niveau de traitement élevé en comparaison au niveau de base sur l'ensemble des complications cardiométaboliques. Cependant, puisque WBC semble être une variable de confusion pour au moins deux troubles métaboliques (obésité et résistance à l'insuline) et que cette variable est exclue du modèle où les estimations des $ATE^{N\geq k}$, $k = 1, \ldots, 4$, sont statistiquement

significatives, ces résultats devraient être pris avec prudence.

Tableau 2.4 Différences de risque (ATEs) pour les facteurs de risque métabolique cumulés ($N = \sum_j Y^j$). *Les estimations ponctuelles sont les moyennes a posteriori obtenues avec* 10000 *réalisations.*

	ATE (Intervalle de crédibilité 95%)	
	Ajusté incluant WBC	Ajusté excluant WBC
$N \geq 1$		
DF/RT	0.16 (-0.04, 0.34)	0.17 (-0.03, 0.35)
DE/RT	0.15 (0.01, 0.30)	0.19 (0.05, 0.33)
$N \geq 2$		
DF/RT	0.11 (-0.06, 0.29)	0.12 (-0.04, 0.30)
DE/RT	0.07 (-0.05, 0.19)	0.13 (0.02, 0.24)
$N \geq 3$		
DF/RT	0.04 (-0.03, 0.14)	0.05 (-0.02, 0.14)
DE/RT	0.03 (-0.02, 0.08)	0.06 (0.01, 0.12)
$N = 4$		
DF/RT	0.00 (-0.00, 0.02)	0.00 (-0.00, 0.01)
DE/RT	0.01 (-0.00, 0.02)	0.01 (0.00, 0.02)

voir table 2.2 pour les notes et légendes.

À la table 2.5, on trouve les rapports de cotes associés aux variables d'ajustement pour chaque trouble métabolique sous les trois modèles étudiés : modèle ajusté complet, modèle ajusté sans WBC, modèle ajusté complet restreint au groupe de survivants ayant reçu DE/RT. En examinant la table, on remarque que, dans les trois modèles, le fait d'être un homme diminue significativement le risque d'être obèse en même temps qu'augmente le risque de souffrir de (pré)-hypertension.

Dans les mêmes modèles, on voit de plus que le temps depuis le diagnostic est associé positivement au risque d'être dyslipidémique.

En ce qui concerne WBC, on rappelle qu'on a constaté que celle-ci était présumément une variable de confusion pour l'estimation de l'effet du traitement sur la résistance à l'insuline. En effet des résultats différents pour le traitement étaient obtenus selon l'inclusion de WBC dans le modèle. Or, l'association remarquée entre WBC et Y^I dans le modèle ajusté complet peut être induite par l'existence de deux chemins causaux (voir figure 2.1), l'un allant de WBC à Y^I (effet direct), et l'autre par l'entremise de la variable TXT utilisée comme proxy unique pour les autres agents chimiothérapeutiques non considérés dans nos analyses. Afin de pouvoir identifier l'effet direct de WBC sur la résistance à l'insuline, il est nécessaire de fixer TXT et CS simultanément. L'analyse stratifiée, obtenue en restreignant le modèle ajusté au groupe de survivants DE/RT, a été réalisée pour atteindre ce but. Essentiellement, la restriction sur le groupe DE/RT, qui est le groupe de survivants à plus haut risque, permet implicitement de fixer les niveaux de CS et de TXT, bloquant ainsi les chemins 2 et 3 du diagramme causal de la figure 2.1. Ainsi, le résultat de l'analyse stratifiée obtenu pour WBC, $OR = 1.02$ (IC à 95 % : $1.01, 1.03$) ou $OR = 1.17$ (IC à 95 % : $1.07, 1.29$) pour un incrément de 10 unités, indique que la concentration de cellules blanches au diagnostic pourrait avoir un impact direct sur la prévalence de la résistance à l'insuline chez les survivants.

Finalement, les figures 2.2 et 2.3 présentent la force de l'association entre WBC et la résistance à l'insuline lorsque WBC est définie comme une variable binaire. Plus précisément, la figure 2.2 montre WBC vs $\log OR$ obtenus en ajustant une suite de régressions logistiques stantards où, pour chacune, WBC a été dichotomisée selon une valeur de seuil croissante fixée entre 3 et 120 ($\times 10^9$/L). La figure 2.3 montre le même type de graphique, mais lorsqu'on restreint l'analyse au groupe DE/RT.

Dans les deux graphiques, on remarque une tendance croissante de $\log(OR)$ ainsi que des valeurs très significatives dépassant 2.00 ($OR > 7.40$) lorsque les seuils de WBC sont proches de 120. Ainsi, ces derniers résultats viennent conforter l'hypothèse de l'effet direct de WBC sur la résistance à l'insuline. Globalement, ces résultats semblent indiquer que la prévalence élevée de complications cardiométaboliques ne serait pas exclusivement due au traitement de la LALe, mais serait également due à la maladie elle-même.

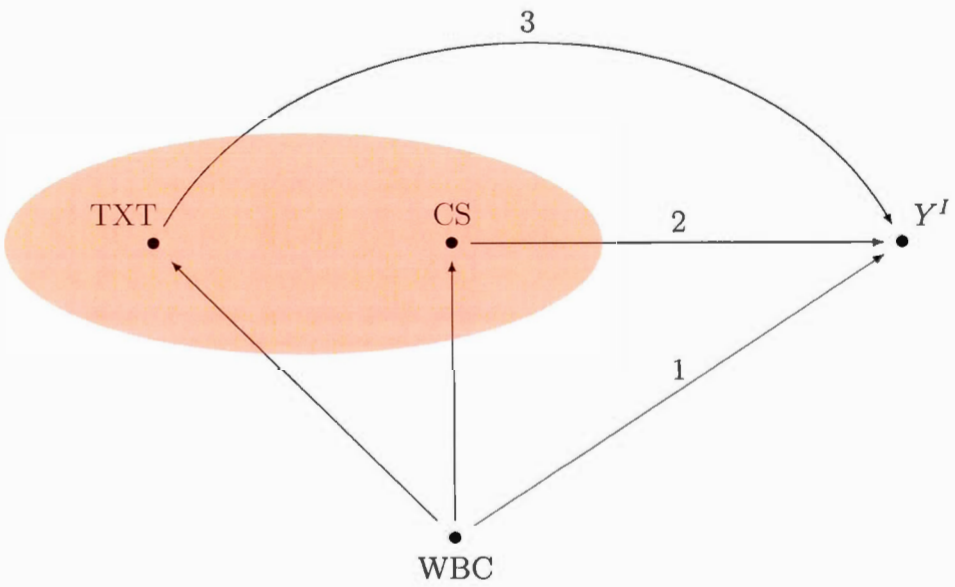

Figure 2.1 Liens causaux potentiels entre les niveaux de doses de corticostéroïdes (CS), d'autres agents chimiothérapeutiques (TXT), les WBC et la résistance à l'insuline (Y^I). Fixer TXT et CS permet d'isoler et estimer la force du lien 1.

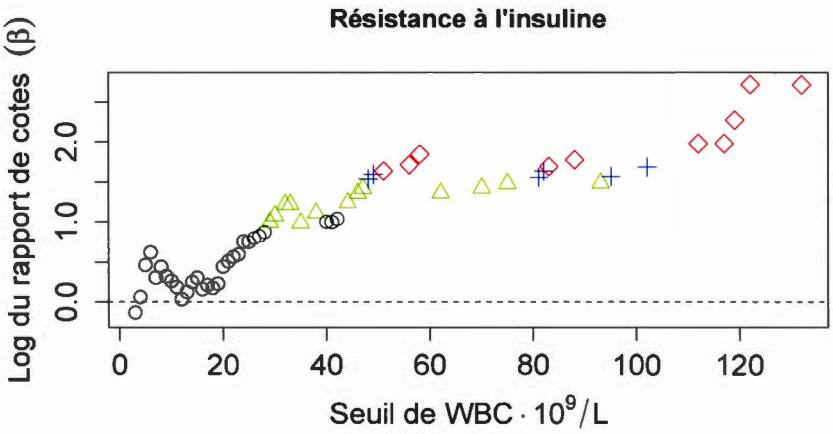

Figure 2.2 Rapports de cotes (sur échelle logarithmique) ajustés associés à une paramétrisation binaire de WBC pour résistance à l'insuline, où la paramétrisation est basée sur des seuils croissants entre 3 et 120. Symboles de signification statistique : cercles pour P (p-value) > 0.05 ; triangles pour $0.01 < P < 0.05$; croix pour $0.005 < P < 0.01$; losanges pour $P < 0.005$.

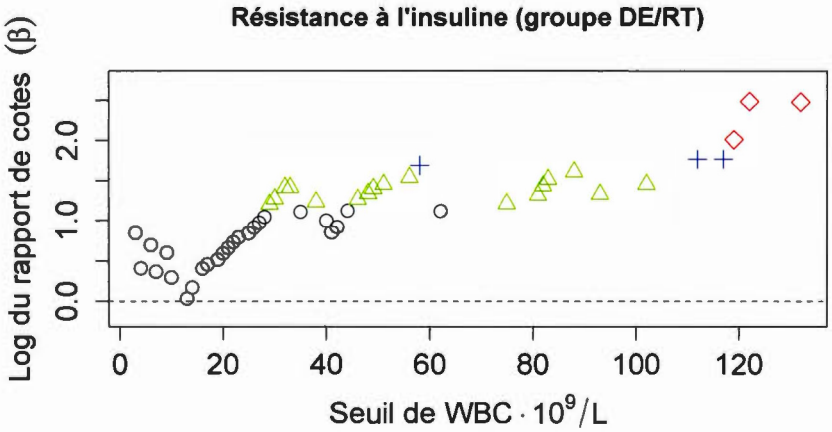

Figure 2.3 Rapports de cotes (sur échelle logarithmique) ajustés associés à des WBC binaires pour résistance à l'insuline. Voir graphique 2.2 pour les symboles de signification statistique

Tableau 2.5 Rapports de cotes (*ORs*) ajustés pour les facteurs de risque métabolique individuels. *Les estimations ponctuelles sont la moyenne/médiane a posteriori obtenues avec* 10000 *réalisations.*

	OR (Intervalle de crédibilité à 95 %)		
	Ajusté incluant WBC	Ajusté sans inclure WBC	Survivants DE/RT seulement
Obésité			
DF/RT †	1.75/1.50 (0.47, 4.37)	1.85/1.59 (0.53, 4.72)	NA
DE/RT †	1.18/1.09 (0.49, 2.40)	1.46/1.36 (0.65, 2.85)	NA
Âge au at diag. (années)	1.02/1.02 (0.93, 1.11)	1.02/1.02 (0.93, 1.11)	1.03/1.03 (0.92, 1.16)
Temps depuis diag. (années)	1.04/1.04 (0.97, 1.12)	1.03/1.03 (0.96, 1.11)	0.98/0.98 (0.89, 1.08)
Homme	0.50/0.47 (0.23, 0.94)*	0.52/0.49 (0.24, 0.99)*	0.42/0.38 (0.13, 0.96) *
WBC ($\times 10^9$/L)	1.01/1.01 (1.00, 1.01)	NA	1.01/1.01 (1.00, 1.01)
Résistance à l'insuline			
DF/RT	1.71/1.39 (0.34, 5.03)	2.04/1.66 (0.43, 6.03)	NA
DE/RT	0.98/0.85 (0.29, 2.46)	2.09/1.85 (0.75, 4.84)	NA
Âge au at diag. (années)	1.02/1.02 (0.91, 1.14)	1.02/1.02 (0.92, 1.14)	0.99/0.99 (0.85, 1.14)
Temps depuis diag. (années)	1.10/1.10 (1.00, 1.21)*	1.08/1.08 (0.99, 1.17)	1.06/1.05 (0.93, 1.20)
Homme	0.84/0.76 (0.30, 1.84)	0.92/0.85 (0.36, 1.92)	1.26/1.06 (0.33, 3.39)
WBC ($\times 10^9$/L)	1.02/1.02 (1.01, 1.03)*	NA	1.02/1.02 (1.01, 1.03) *
(Pré)-hypertension			
DF/RT	1.29/0.93 (0.15, 4.69)	1.25/0.89 (0.14, 4.51)	NA
DE/RT	3.03/2.40 (0.71, 8.85)	3.03/2.42 (0.74, 9.07)	NA
Âge au at diag. (années)	1.01/1.01 (0.88, 1.14)	1.00/1.00 (0.88, 1.13)	1.02/1.02 (0.86, 1.19)
Temps depuis diag. (années)	0.95/0.95 (0.84, 1.07)	0.95/0.95 (0.84, 1.07)	0.89/0.89 (0.75, 1.03)
Homme	6.04/4.89 (1.70, 17.2)*	6.06/4.91 (1.70, 17.1)*	9.38/6.58 (1.72, 33.5) *
WBC ($\times 10^9$/L)	1.00/1.00 (0.99, 1.01)	NA	1.00/1.00 (0.98, 1.01)
Dyslipidémie			
DF/RT	2.70/2.36 (0.84, 6.65)	2.63/2.30 (0.83, 6.40)	NA
DE/RT	2.70/2.49 (1.16, 5.48)*	2.51/2.34 (1.15, 4.81)*	NA
Âge au at diag. (années)	1.06/1.06 (0.97, 1.15)	1.06/1.06 (0.97, 1.15)	1.09/1.09 (0.98, 1.22)
Temps depuis diag. (années)	1.08/1.08 (1.01, 1.16)*	1.08/1.08 (1.01, 1.16)*	1.13/1.13 (1.02, 1.25) *
Homme	1.18/1.11 (0.58, 2.14)	1.17/1.11 (0.57, 2.16)	0.78/0.70 (0.28, 1.75)
WBC ($\times 10^9$/L)	1.00/1.00 (0.99, 1.01)	NA	1.00/1.00 (0.99, 1.01)

* IC ne contient pas la valeur de référence (résultats significatifs).
voir table 2.2 pour d'autres notes et légendes.

2.4.2 Résultats de la régression logistique standard

Pour fins de comparaison, on présente, à la table 2.6, les estimations ponctuelles des ORs ainsi que des intervalles de confiance à 95% pour ceux-ci. On constate que, à l'exception de trois cas, les valeurs des estimations par maximum de vraisemblance des ORs se trouvent toujours entre la moyenne et la médiane de leurs distributions *a posteriori* (voir table 2.2). On remarque aussi que les intervalles de confiance sont très semblables, mais toujours un peu plus larges, aux intervalles de crédibilité présentés à la table 2.2. Ainsi, ces résultats viennent conforter les affirmations faites précédemment concernant l'effet du traitement sur chaque réponse cardiométabolique individuellement et plus globalement, l'interprétation marginale du modèle t-link.

Tableau 2.6 Rapports de cotes (ORs) bruts et ajustés associés au traitement pour les facteurs de risque cardiométabolique individuels. *Estimations ponctuelles et intervalles de confiance obtenues par maximum de vraisemblance.*

	OR (Intervalle de confiance 95%)		
	Brut	Ajusté incluant WBC	Ajusté excluant WBC
Obésité			
DF/RT	1.37 (0.49, 3.66)	1.56 (0.51, 4.69)	1.70 (0.56, 5.03)
DE/RT	1.34 (0.68, 2.72)	1.09 (0.48, 2.46)	1.38 (0.66, 2.93)
Résistance à l'insuline			
DF/RT	2.07 (0.57, 6.99)	1.40 (0.34, 5.41)	1.84 (0.46, 6.88)
DE/RT	1.97 (0.82, 5.12)	0.78 (0.25, 2.43)	1.73 (0.67, 4.73)
(Pré)-hypertension			
DF/RT	1.52 (0.20, 8.37)	0.97 (0.12, 5.86)	0.96 (0.12, 5.77)
DE/RT	2.75 (0.91, 10.23)	2.67 (0.71, 11.55)	2.59 (0.74, 10.63)
Dyslipidémie			
DF/RT	3.18 (1.21, 8.49)	2.45 (0.86, 7.08)	2.40 (0.85, 6.89)
DE/RT	2.96 (1.50, 6.02)	2.54 (1.16, 5.63)	2.39 (1.15, 5.06)

voir table 2.2 pour les notes et légendes.

CONCLUSION

L'étude présentée dans ce mémoire avait comme principal objectif de contribuer à une meilleure compréhension du rôle du traitement dans le développement des complications cardiométaboliques chez les survivants de la LALe. L'accès aux données de la cohorte PÉTALE nous a permis de mener des analyses qui, pour la première fois selon nos connaissances, visaient à estimer l'effet combiné des doses reçues de CS et l'exposition à la RT sur le risque de développer certains EIL cardiométaboliques chez des survivants de la LALe. Par ailleurs, puisque les quatre complications étudiées : obésité, résistance à l'insuline, (pré)-hypertension et dyslipidémie, forment un groupe cohésif de réponses, on a décidé d'adopter une approche multivariée pour l'analyse des données. Plus spécifiquement, notre analyse statistique s'est basée sur le modèle de régression logistique multivarié bayésien proposé par O'Brien et Dunson (2004) jumelé au cadre bayésien de Hund *et al.* (2015) pour l'estimation des effets de l'exposition.

Grâce à cette stratégie de modélisation et d'inférence, on a pu estimer l'effet du traitement sur l'ensemble des réponses cardiométaboliques de façon marginale et conjointe, à l'aide de deux types de mesures d'effet facilement interprétables dans le domaine biomédical. Ainsi, cette approche nous a donné une description plus précise de l'influence des niveaux de traitement sur les réponses comparativement à une série d'analyses univariées, en plus de mieux contrôler le risque de commettre une erreur de première espèce et d'accroître la puissance statistique (Tabachnik et Fidell, 2013; Yang et Wang, 2012). En fait, la puissance statistique était une source de préoccupation dans notre étude étant donnée la taille modeste de la cohorte ($n = 180$). En ce qui concerne les simulations des lois *a posteriori* au moyen

de l'algorithme MCMC proposé par O'Brien et Dunson (2004), on a constaté une forte dépendance entre réalisations consécutives de la matrice de corrélation $\boldsymbol{R}$, et donc une efficacité computationnelle moindre que prévue. Faute d'analyses spécifiques qui explorent cette problématique, on pense que la dimensionnalité de la matrice $\boldsymbol{R}$ pourrait être étroitement liée à la difficulté d'explorer efficacement l'espace d'état. De ce fait, il semble pertinent de mentionner que l'analyse statistique effectuée a une complexité plus grande que celles des applications présentées dans O'Brien et Dunson (2004) et Hund *et al.* (2015). Plus précisément, O'Brien et Dunson illustrent la méthode au moyen d'un cas d'étude où interviennent 6 variables réponses et seulement 7 coefficients de régression (communs pour toutes les réponses), tandis que Hund *et al.* présente un cas où interviennent 3 variables réponses et entre 4 et 6 coefficients de régression selon la réponse. Ainsi, notre cas d'étude avec 4 variables réponses pour un total de 28 coefficients à estimer est clairement plus complexe.

Les résultats des analyses effectuées montrent que, chez les jeunes survivants de la LALe, recevoir le niveau de traitement le plus élevé (DE/RT) augmente de 0.19 le risque de présenter de la dyslipidémie comparativement au traitement de base (DF/non RT) (table 2.3). De plus, ce même groupe de survivants (DE/RT) a un risque plus élevé de subir au moins une complication cardiométabolique : différence de risque avec le groupe de référence de 0.15 (table 2.4). Bien que pas statistiquement significatif au seuil 5%, le risque en excès de présenter de la dyslipidémie seulement et aucune autre complication est assez grand (0.10), ce qui explique la magnitude des deux risques en excès mentionnés ci-dessus. L'approche de modélisation multivariée nous a aussi permis de constater que le niveau traitement le plus élevé augmente légèrement, par rapport au traitement de base, la probabilité d'être atteint de (pré)-hypertension et de dyslipidémie simultanément, sans obésité ni résistance à l'insuline (risque en excès de 0.02). Par ailleurs, nous

avons trouvé des différences de risque pour trois ou quatre complications cumulées petites et statistiquement non significatives, ce qui nous laisse penser que le traitement combiné des doses de CS et de RT n'est pas lié à un risque accru de présenter plusieurs complications cardiométaboliques simultanément. On souligne aussi que les ORs et ATEs pour la dyslipidémie associés au traitement DF/RT (versus DF/ non RT), bien que pas significatifs, sont similaires en magnitude à ceux associés au niveau DE/RT. Ceci semble indiquer que c'est plutôt la RT, et non pas les doses de CS, le facteur déterminant qui expliquerait l'impact du traitement sur le risque de présenter la dyslipidémie. De fait, dans l'étude SJLIFE sur la survie en cancer pédiatrique (Nottage, *et al.*, 2014) aucune association cliniquement significative entre les agents chimiothérapeutiques et les complications cardiométaboliques étudiées n'a été trouvée.

Finalement, nos analyses suggèrent que le traitement ne serait pas le principal facteur explicatif de la haute prévalence de la résistance à l'insuline chez les jeunes survivants de la LALe. Il semble, plutôt, que l'élément clé soit la sévérité de la maladie, mesurée par la concentration de cellules blanches au diagnostic (WBC). En fait, l'augmentation des WBC est un biomarqueur du niveau d'activation du système immunitaire et des processus inflammatoires qui, à leur tour, semblent être des facteurs contributifs au développement de la résistance à l'insuline (Olefsky et Glass, 2010). Bien que des liens entre un nombre élevé de WBC et le diabète de type 2 ont déjà été trouvés dans différentes populations (Vozarova *et al.*, 2002; Yoshimura *et al.*, 2015), à notre connaissance, la présente étude est la première à enquêter sur une possible association entre la sévérité (progression) de la LALe au diagnostic et la résistance à l'insuline chez les survivants. Si d'autres études valident ces résultats, ceux-ci pourraient éventuellement mener à un meilleur suivi personnalisé des survivants de LALe en fonction de leur risque individuel.

ANNEXE A

VALIDATION INFORMATIQUE DU MODÈLE

Nous présentons dans cette annexe des résultats portant sur l'analyse de deux jeux de données simulés, pour fins de validation du code informatique implémenté pour le modèle de O'Brien et Dunson (2004).

Pour les deux jeux de données, on a simulé $n = 10000$ réalisations des vecteurs de réponses binaires $\boldsymbol{Y}_i = (Y_i^1, Y_i^2, Y_i^3)$, $i = 1, \ldots, n$, au moyen de la loi logistique multivariée de O'Brien et Dunson (2004). Pour le premier jeu de données, plus spécifiquement, chaque vecteur $\boldsymbol{Y}_i$ a été obtenu en utilisant la relation :

$$Y_i^j = \mathbb{1}(Z_i^j > 0) \quad j = 1, 2, 3$$

$$\boldsymbol{Z}_i = (Z_i^1, Z_i^2, Z_i^3) \sim \mathcal{L}_p(\boldsymbol{z}_i | \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{R}, \nu),$$

$$\boldsymbol{X}_i\boldsymbol{\beta} = \text{diag}(\boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^3)((\boldsymbol{\beta}^1)', (\boldsymbol{\beta}^2)', (\boldsymbol{\beta}^3)')', \; \boldsymbol{X}_i^j\boldsymbol{\beta}^j = \beta_0^j + X_{1,i}\beta_1^j + X_{2,i}\beta_2^j,$$

avec les paramètres du modèle fixés aux valeurs suivantes : $\boldsymbol{\beta}^1 = (\beta_0^1, \beta_1^1, \beta_2^1) = (2, -1.5, 0.7)$, $\boldsymbol{\beta}^2 = (\beta_0^2, \beta_1^2, \beta_2^2) = (-3, 2, 1.7)$, $\boldsymbol{\beta}^3 = (\beta_0^3, \beta_1^3, \beta_2^3) = (-2, 3.2, 0.3)$, $\nu = 7.3$ et

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

Les covariables $X_{1,i}$, $i = 1, \ldots, n$, ont été générées indépendamment au moyen

d'une loi binomiale de paramètre ($p = 0.3$), tandis que les $X_{2,i}$ sont des covariables continues issues d'une loi normale centrée réduite. Les vecteurs $\boldsymbol{Z}_i$ ont été obtenus en simulant des vecteurs Student et en appliquant la transformation décrite au chapitre 1.2.3. Ce faisant, cette validation a également comme but d'évaluer l'approximation du modèle t-link pour l'analyse de données logistiques multivariées.

Pour simuler le deuxième jeux de données, on a suivi la même procédure en fixant les mêmes valeurs pour les coefficients de régression mais en modifiant celles des coefficients de corrélation de 0.5 (pour tous) à :

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0.8 \\ 0.5 & 1 & 0.3 \\ 0.8 & 0.3 & 1 \end{pmatrix}.$$

Les résumés des lois *a posteriori* estimées pour les coefficients de régression et de corrélation au moyen d'un échantillon de taille 10000 sont présentés dans les tableaux A.1 à A.4.

Dans les figures A.1 à A.4 on présente également les densités estimées pour chaque paramètre du modèle. On remarque que, pour tous les paramètres, les moyennes empiriques sont assez proches des vraies valeurs. Cependant, dans certains cas, les vraies valeurs se trouvent près des bornes des intervalles de crédibilité. Quelques raisons peuvent être avancées pour expliquer ceci. Premièrement, le modèle *t-link* utilisé pour l'inférence n'est pas le même que le modèle original de O'Brien et Dunson (2004), bien qu'il en soit une très bonne approximation. De plus, les lois *a priori* des paramètres peuvent aussi avoir une légère influence sur les résultats obtenus malgré la relativement grande taille ($n = 10000$) des jeux des données.

Tableau A.1 Jeu de données 1 : résumés des lois *a posteriori* des coefficients de régression.

Covariables		Vraies valeurs	Moyenne (Intervalle de crédibilité 95%)
Ordonnée à l'origine	β_0^1	2	1.94 (1.87, 2.02)
X_1	β_1^1	-1.5	-1.53 (-1.63, -1.43)
X_2	β_2^1	0.7	0.68 (0.63, 0.74)
Ordonnée à l'origine	β_0^2	-3	-3.00 (-3.12, -2.88)
X_1	β_1^2	2	1.96 (1.82, 2.09)
X_2	β_2^2	1.7	1.64 (1.58, 1.72)
Ordonnée à l'origine	β_0^3	-2	-1.96 (-2.03, -1.89)
X_1	β_1^3	3.2	3.15 (3.04, 3.26)
X_2	β_2^3	0.3	0.28 (0.22, 0.33)

Tableau A.2 Jeu de données 1 : résumés des lois *a posteriori* des coefficients de corrélation. Vraies valeurs : 0.5 pour les trois coefficients.

	Moyenne (Intervalle de crédibilité 95%)		
	Y^1	Y^2	Y^3
Y^1	1.00 (1.00, 1.00)	0.50 (0.45, 0.55)	0.53 (0.49, 0.57)
Y^2		1.00 (1.00, 1.00)	0.48 (0.44, 0.52)
Y^3			1.00 (1.00, 1.00)

Tableau A.3 Jeu de données 2 : résumés des lois *a posteriori* des coefficients de régression.

Covariables		Vraies valeurs	Moyenne (Intervalle de crédibilité 95%)
Ordonnée à l'origine	β_0^1	2	1.93 (1.86, 2.00)
X_1	β_1^1	-1.5	-1.47 (-1.57, -1.36)
X_2	β_2^1	0.7	0.71 (0.65, 0.76)
Ordonnée à l'origine	β_0^2	-3	-2.95 (-3.07, -2.84)
X_1	β_1^2	2	2.02 (1.88, 2.15)
X_2	β_2^2	1.7	1.66 (1.57, 1.75)
Ordonnée à l'origine	β_0^3	-2	-1.95 (-2.02, -1.88)
X_1	β_1^3	3.2	3.09 (2.98, 3.20)
X_2	β_2^3	0.3	0.28 (0.23, 0.33)

Tableau A.4 Jeu de données 2 : résumés des lois *a posteriori* des coefficients de corrélation. Vraies valeurs : 0.5, 0.8 et 0.3.

	Moyenne (Intervalle de crédibilité 95%)		
	Y^1	Y^2	Y^3
Y^1	1.00 (1.00, 1.00)	0.50 (0.45, 0.55)	0.78 (0.75, 0.82)
Y^2		1.00 (1.00, 1.00)	0.30 (0.25, 0.34)
Y^3			1.00 (1.00, 1.00)

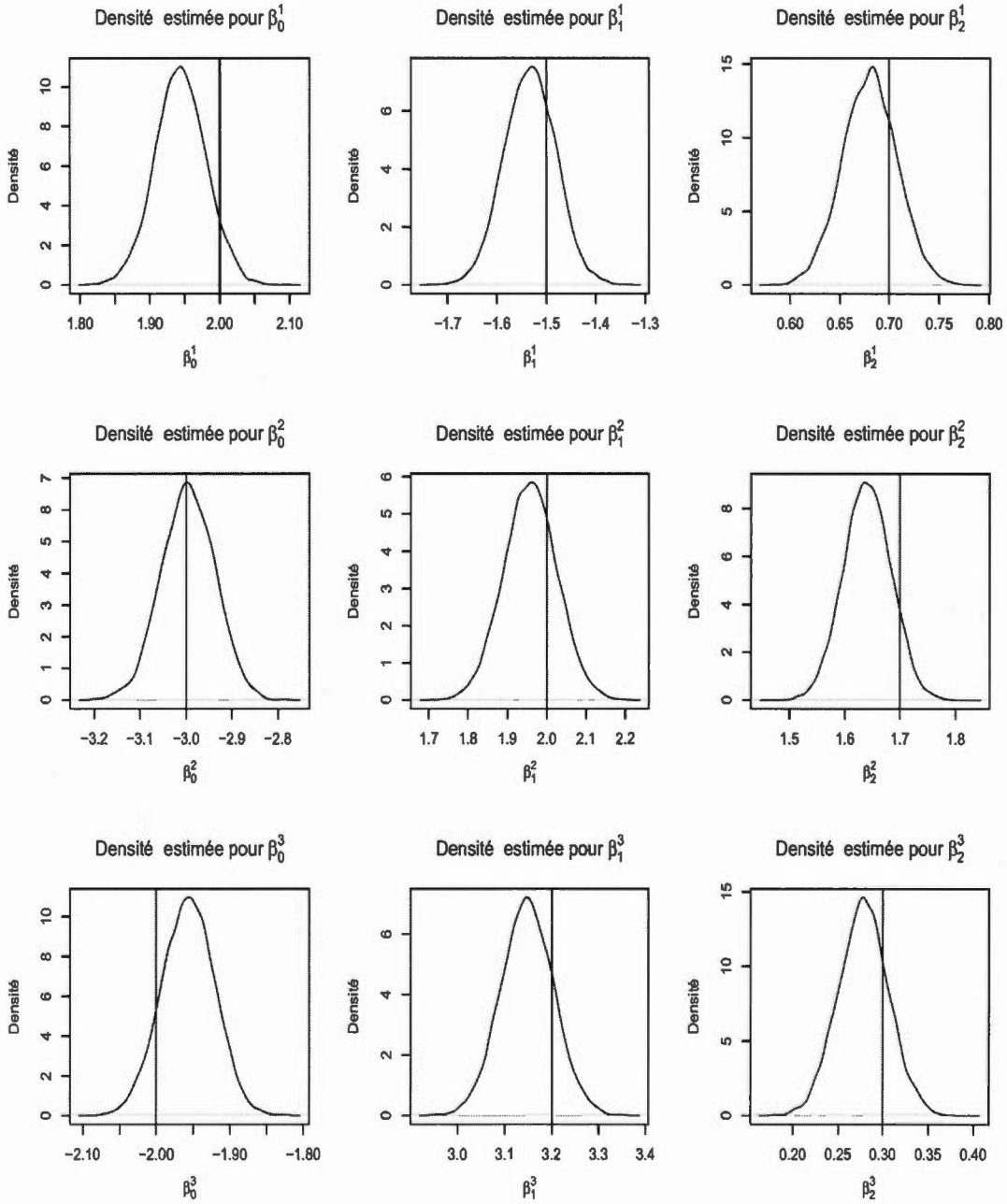

Figure A.1 Jeux de données 1 : densités estimées des lois *a posteriori* des coefficients de régression. La barre verticale indique la vraie valeur du coefficient.

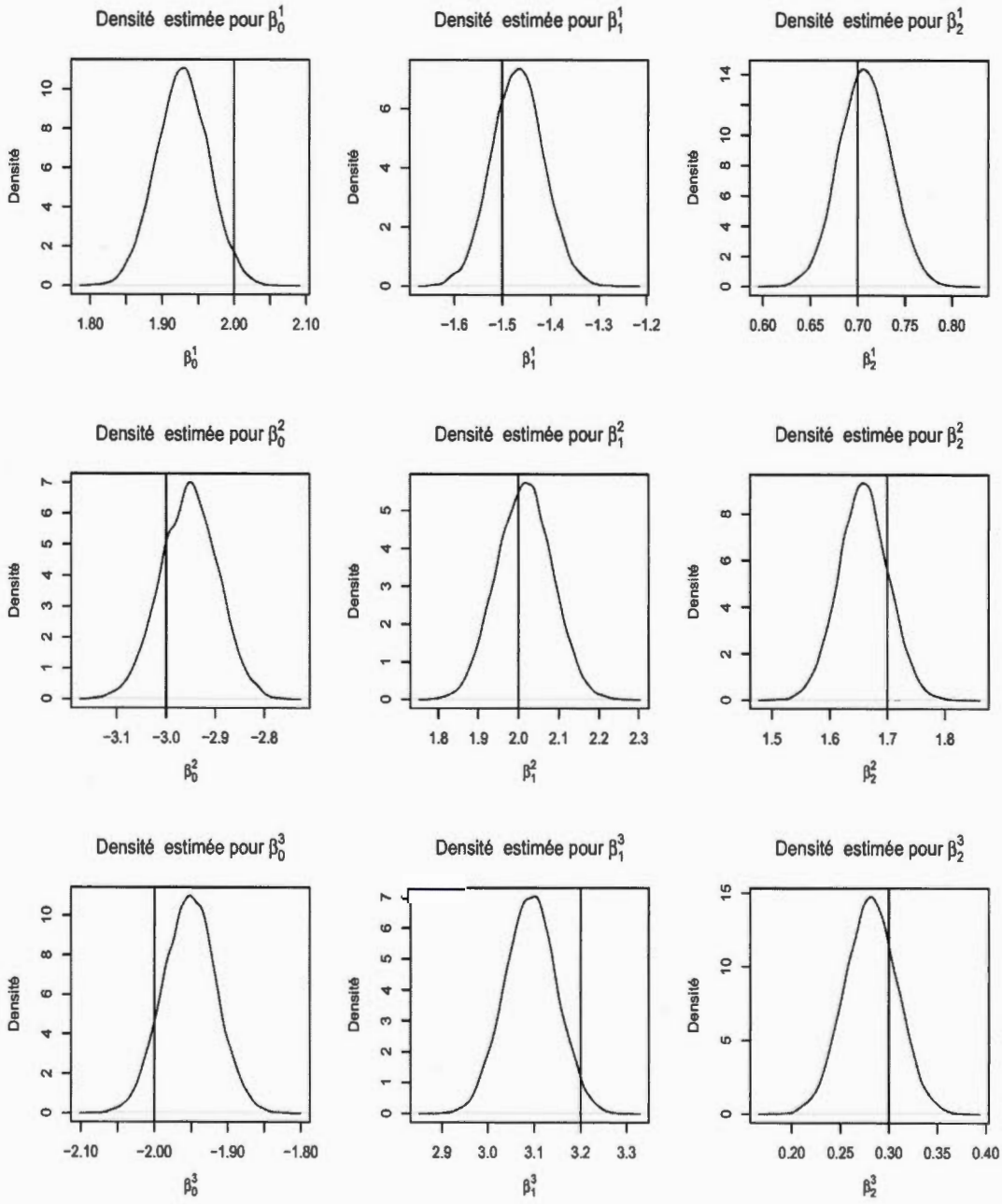

Figure A.2 Jeux de données 2 : densités estimées des lois *a posteriori* des coefficients de régression. La barre verticale indique la vraie valeur du coefficient.

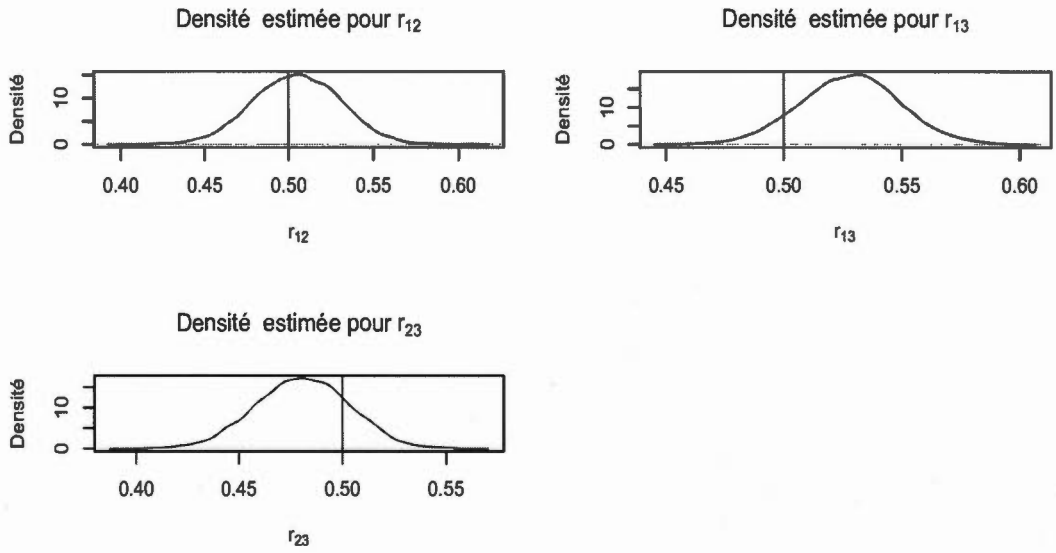

Figure A.3 Jeux de données 1 : densités estimées des lois *a posteriori* des coefficients de corrélation. La barre verticale indique la vraie valeur du coefficient.

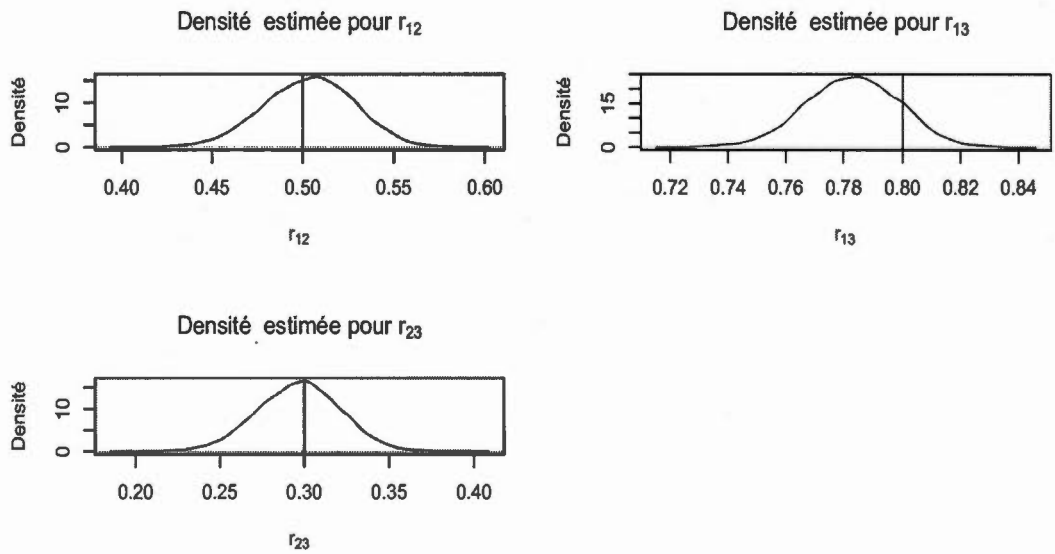

Figure A.4 Jeux de données 2 : densités estimées des lois *a posteriori* des coefficients de corrélation. La barre verticale indique la vraie valeur du coefficient.

ANNEXE B

DIAGNOSTIC DE CONVERGENCE

L'analyse de la convergence vers la loi *a posteriori* des chaînes de Markov générées au moyen de l'algorithme MCMC (appliqué aux données de cette étude) a été effectuée par inspection visuelle des distributions des réalisations de trois chaînes différentes. Le modèle choisi pour faire ce diagnostic a été le modèle complètement ajusté, c'est-à-dire le modèle incluant toutes les covariables confondantes (4) dans les fonctions de régression, en plus des variables d'exposition. La description mathématique des relations entre ces variables sous le modèle *t-link* sont présentées à la section 2.3.1. On souligne que, dans le modèle complet, chaque sous-vecteur $\boldsymbol{\beta}^j$, $j \in \{O, I, H, D\}$, du vecteur des coefficients de régression $\boldsymbol{\beta}$ est de dimension sept, $\boldsymbol{\beta}^j = (\beta_0^j, \beta_1^j, \beta_2^j, \ldots, \beta_6^j)$, et que les coefficients β_1^j et β_2^j correspondent aux niveaux de traitement DF/RT et DE/RT respectivement. Ainsi, une première chaîne a été générée selon les conditions initiales 'init1' : $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}^O, \boldsymbol{\beta}^I, \boldsymbol{\beta}^H, \boldsymbol{\beta}^D)^{(0)} = (0, 0, \ldots, 0)$ et $r_{ij}^{(0)} = 0, i \neq j$ ($\boldsymbol{R}^{(0)} = \mathbf{1}$), une seconde avec les conditions initiales 'init2' : $\boldsymbol{\beta}^{(0)} = (10, 10, \ldots, 10)$ et $r_{ij}^{(0)} = 0, i \neq j$ ($\boldsymbol{R}^{(0)} = \mathbf{1}$) et une dernière avec 'init3' : $\boldsymbol{\beta}^{(0)} = (-5, -5, \ldots, -5)$ et $r_{ij}^{(0)} = 0.9, i \neq j$. Dans les trois cas, les réalisations utilisées pour le diagnostic de convergence ont été obtenues en gardant seulement toutes les 100 itérations ('thinning'= 100) d'une simulation de 1050000 itérations de laquelle on a écarté les premières 50000 valeurs ('burn-in' = 50000).

Dans les figures B.1–B.5 on présente des diagrammes en boîtes pour les coefficients de régression associés aux variables de traitement et pour les coefficients de corrélation. On constate que les diagrammes associés à un même paramètre sont sensiblement pareils à travers 5 groupes de 2000 itérations et pour des conditions initiales très différentes ; ceci est donc un bon indice en ce qui concerne la stabilité des chaînes et leur convergence vers la loi cible. Des résultats très similaires ont été obtenus pour les autres paramètres du modèle n'apparaissant pas dans ces figures. Enfin, on souligne que, en tenant compte de la forte similitude (après 'burn-in') entre des chaînes selon différentes conditions initiales et les calculs assez gourmands nécessaires à l'obtention des différences de risque multiple (ATE), les analyses de la section 2.3 ont été faites en simulant une seule chaîne selon la condition initiale 'init1' tel qu'il est décrit à cette même section.

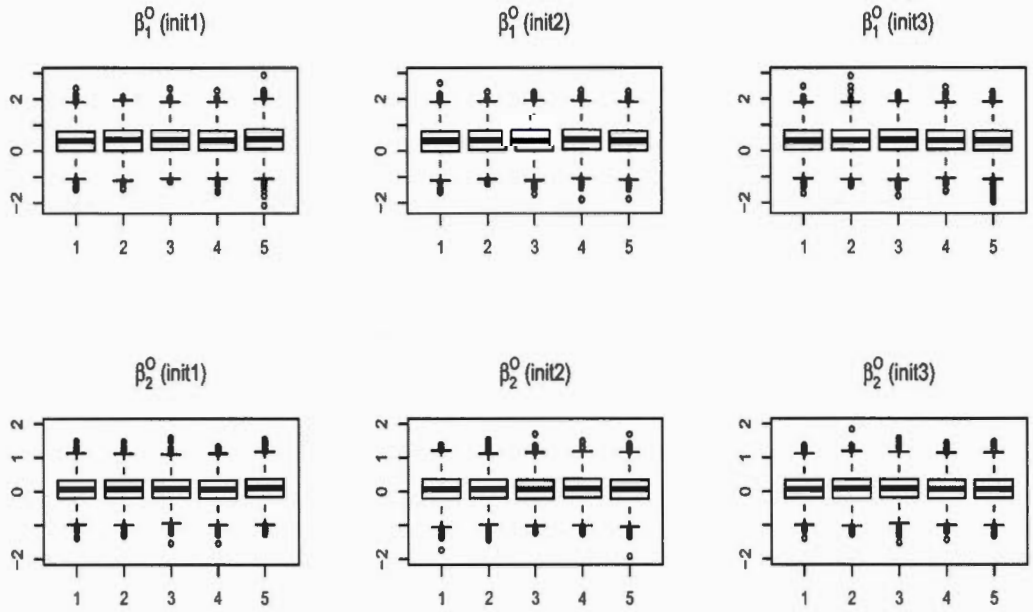

Figure B.1 Diagrammes en boîte pour les coefficients de régression β_1^O, β_2^O par groupe de 2000 itérations consécutives selon les trois conditions initiales.

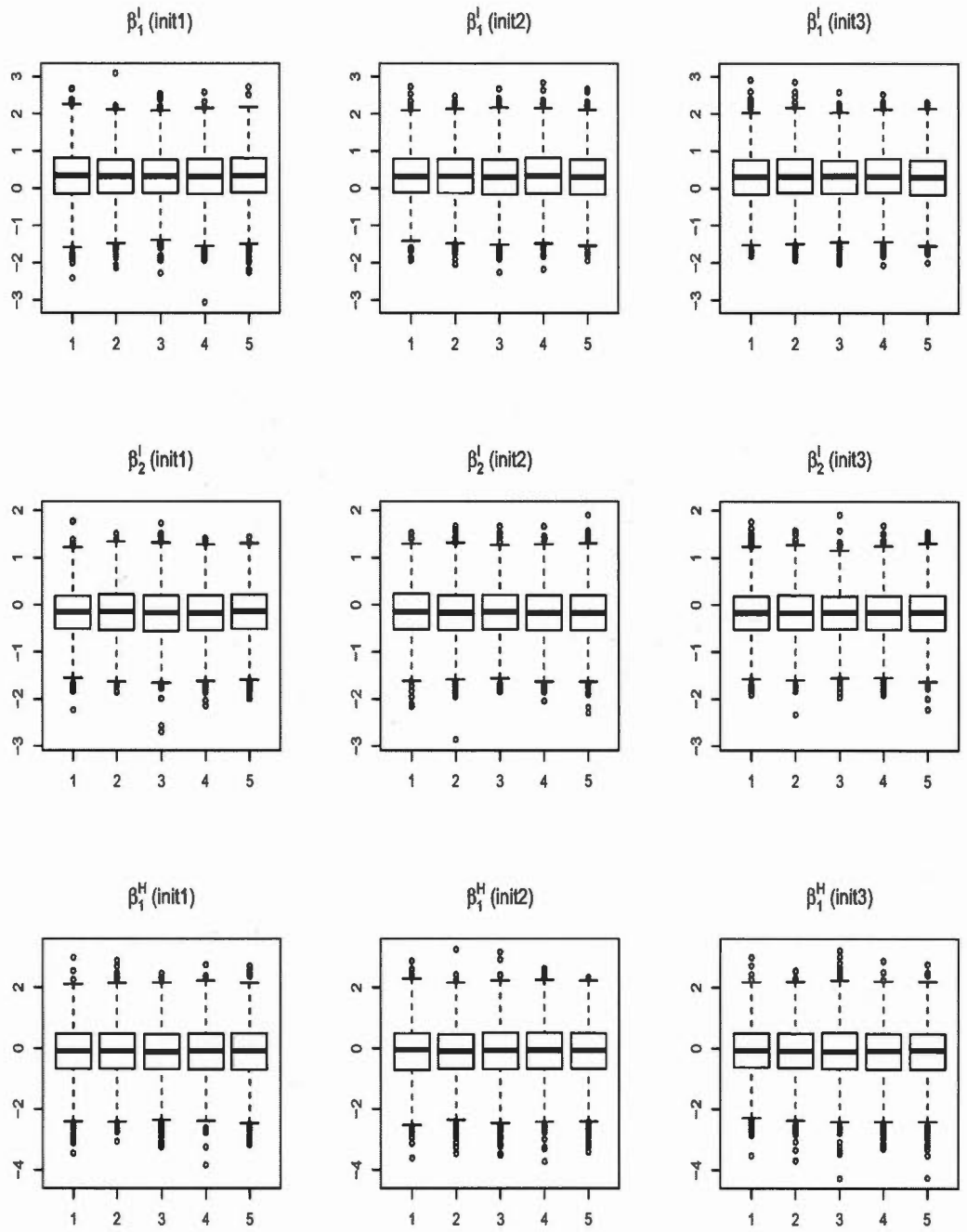

Figure B.2 Diagrammes en boîte pour les coefficients de régression β_1^I, β_2^I, β_1^H par groupe de 2000 itérations consécutives selon les trois conditions initiales.

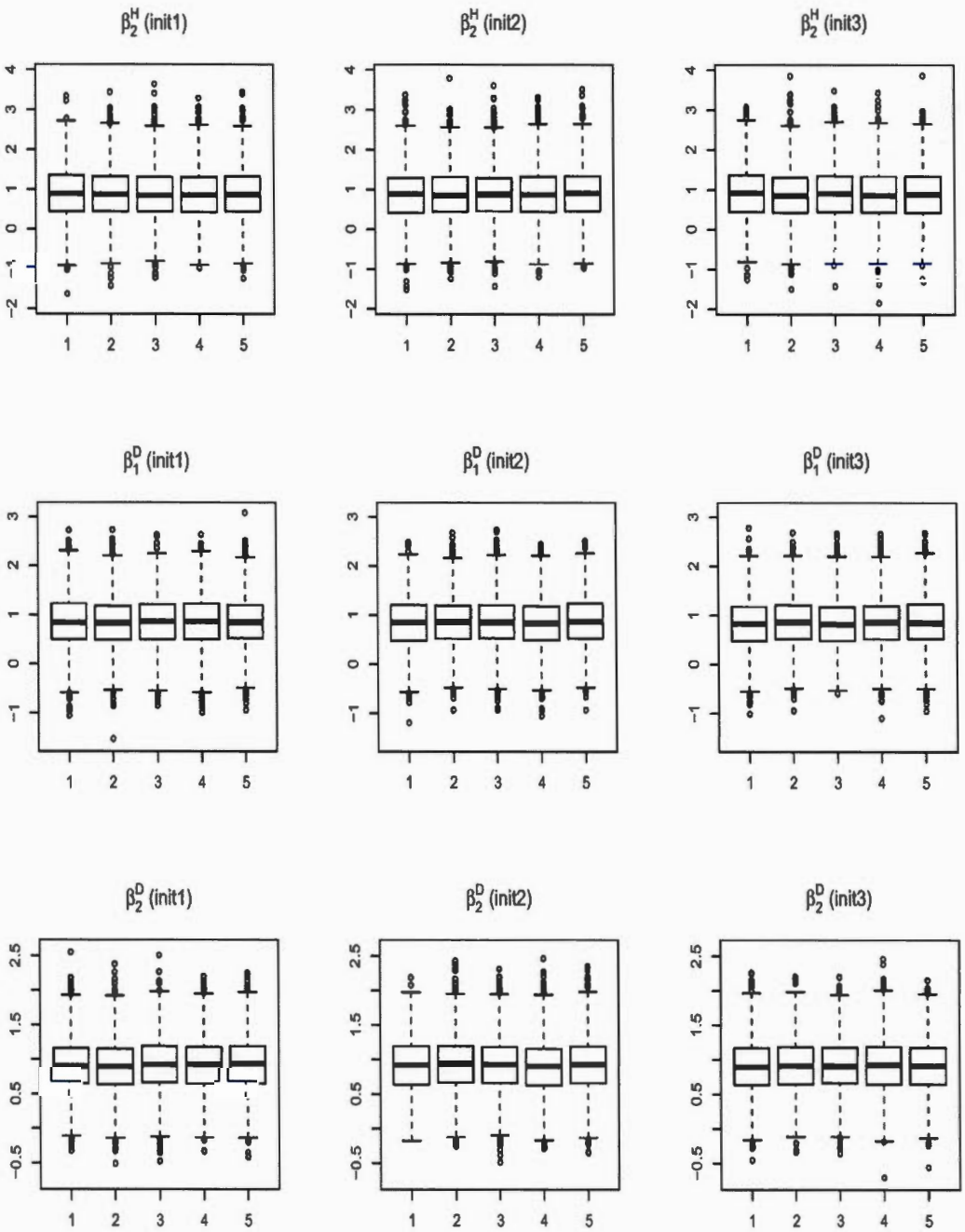

Figure B.3 Diagrammes en boîte pour les coefficients de régression β_2^H, β_1^D, β_2^D par groupe de 2000 itérations consécutives selon les trois conditions initiales.

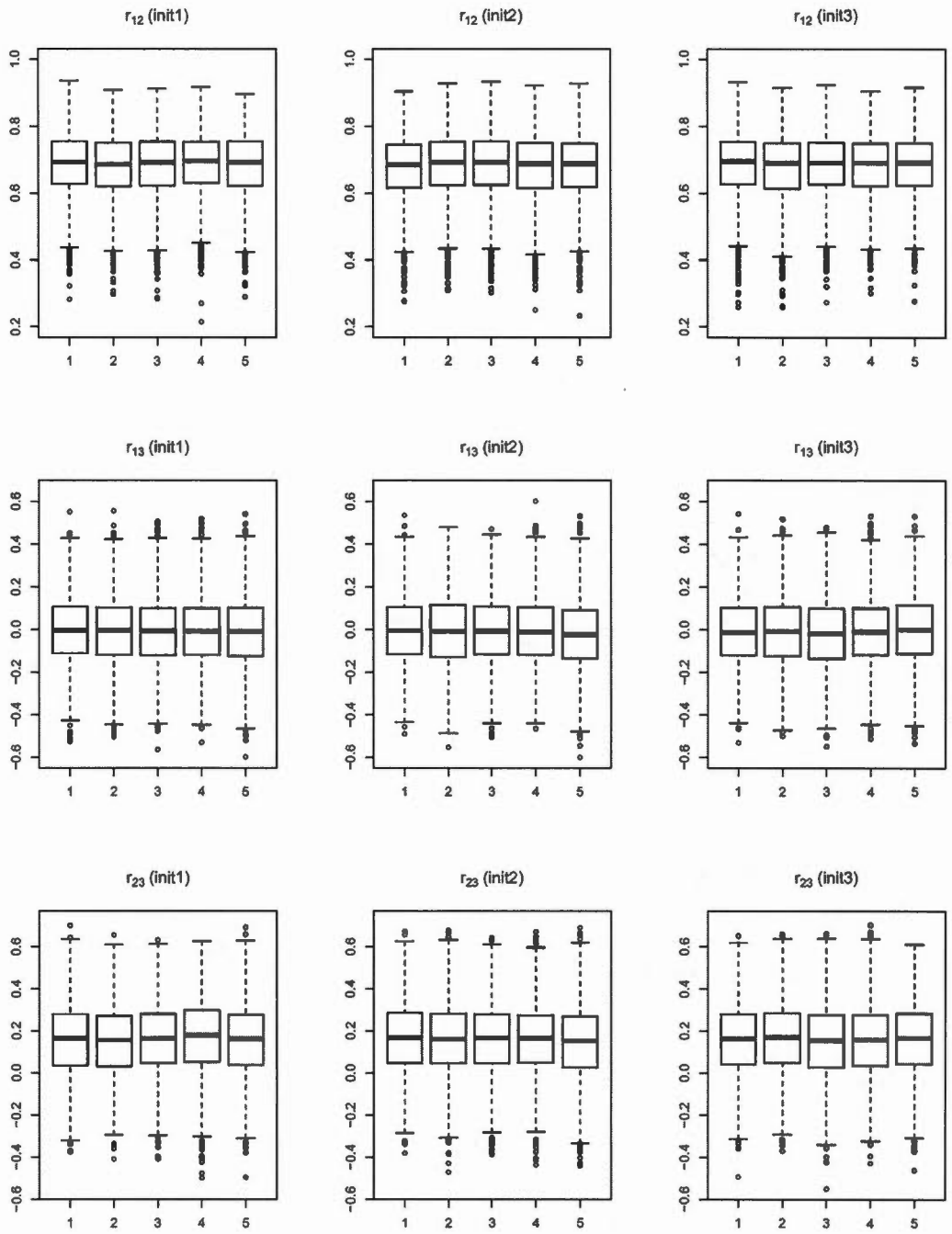

Figure B.4 Diagrammes en boîte pour les coefficients de corrélation r_{12}, r_{13}, r_{23} par groupe de 2000 itérations consécutives selon les trois conditions initiales.

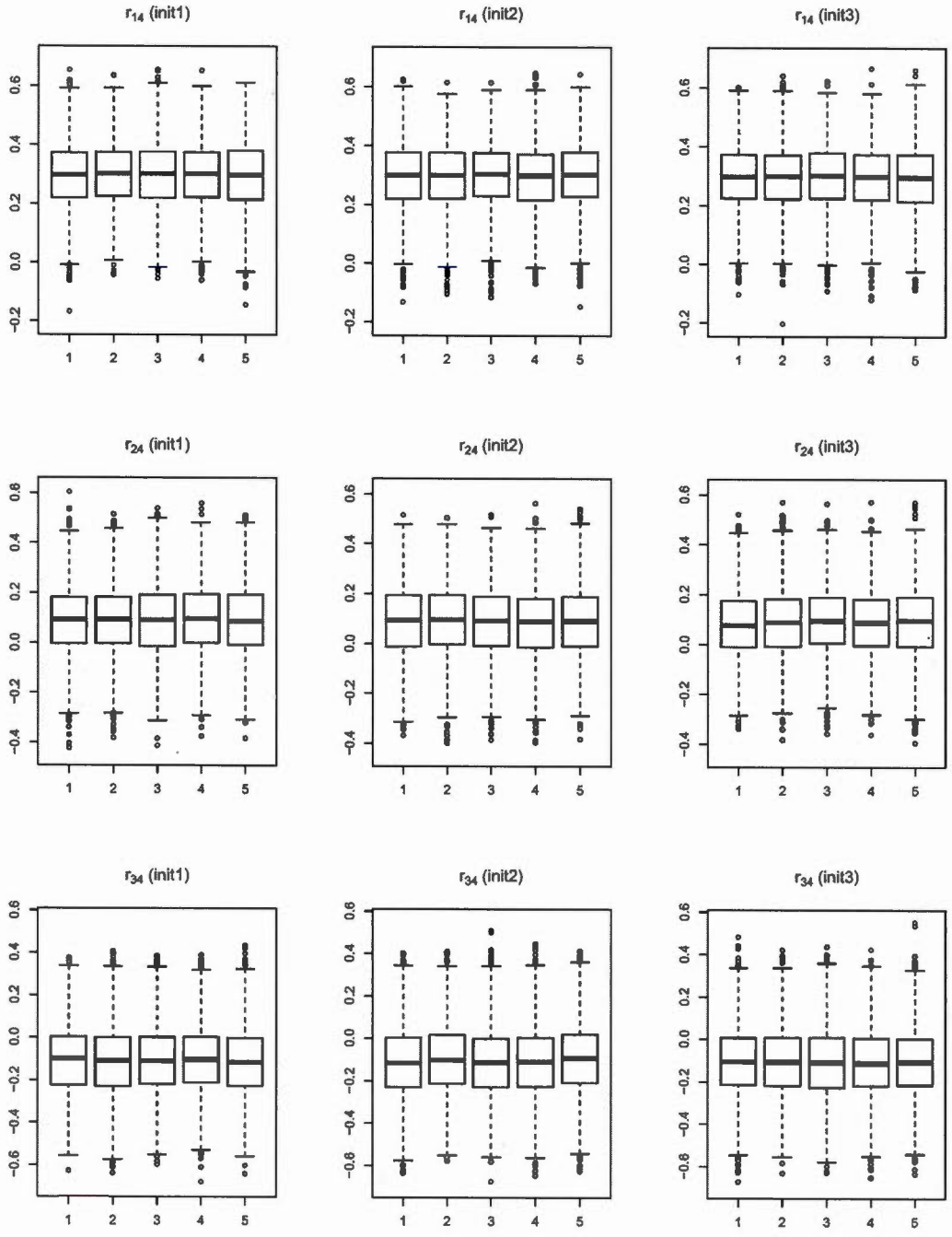

Figure B.5 Diagrammes en boîte pour les coefficients de corrélation r_{14}, r_{24}, r_{34} par groupe de 2000 itérations consécutives selon les conditions initiales.

Ci-dessous on présente les graphiques d'autocorrélation pour les réalisations de deux paramètres (β_2^1, r_{14}) du modèle selon les conditions initiales 'init1'.

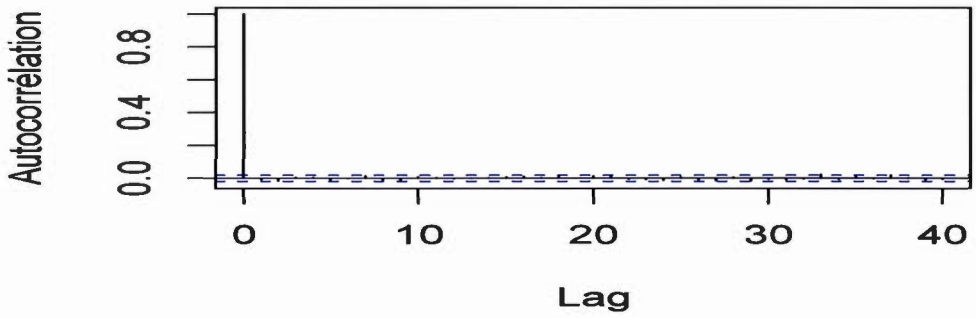

Figure B.6 Graphique d'autocorrélation pour β_2^1

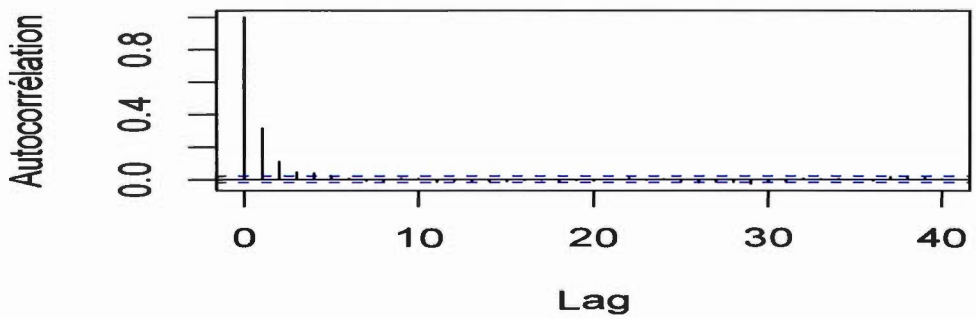

Figure B.7 Graphique d'autocorrélation pour r_{14}

On remarque une faible autocorrélation pour le coefficient de régression β_2^1 tandis que pour le coefficient de corrélation celle-ci est un peu plus élevée. On souligne aussi que, pour tous les autres coefficients de régression et de corrélation, on a obtenu des graphiques d'autocorrélation très similaires à ceux de β_2^1 et r_{14}, respectivement.

Finalement, dans le graphique ci-dessous on peut visualiser la force des corrélations croisées entre les paramètres. On remarque que, dans la plus part de cas, ces corrélations sont faibles.

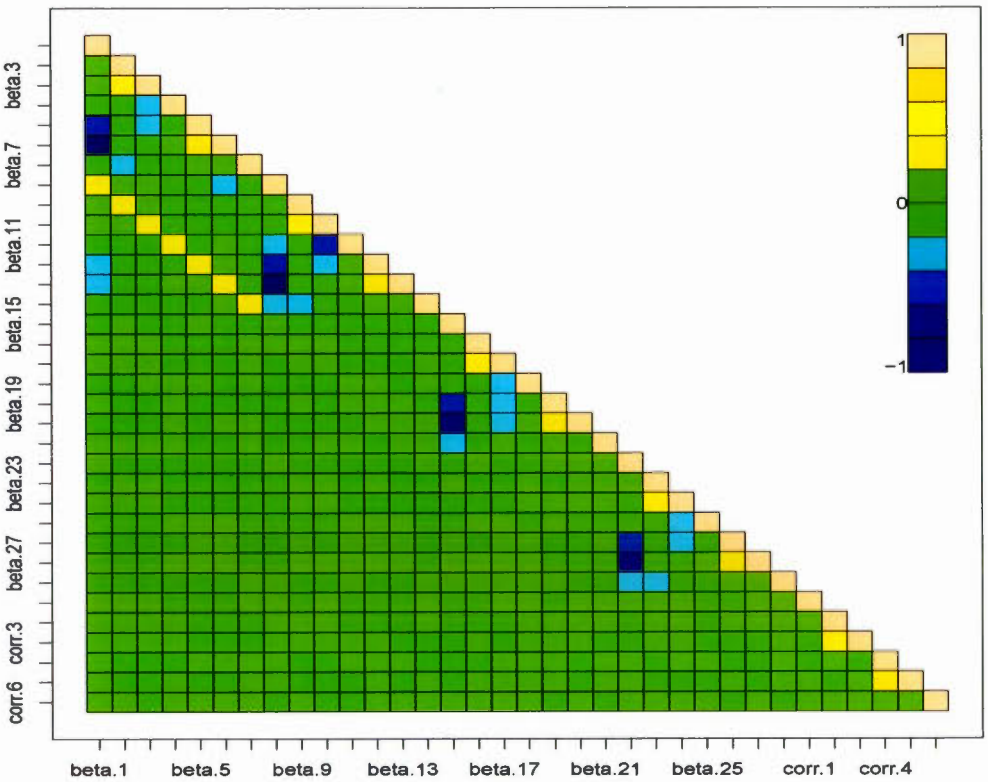

Figure B.8 Graphique de corrélations croisées

ANNEXE C

CODE INFORMATIQUE

```
# Validation informatique du modele

n ←10000
k ←3
beta1 ← c(2,−1.5,0.7)
beta2 ← c(−3,2,1.7)
beta3 ← c(−2,3.2,0.3)
c ← length(beta1)

R ← diag(k)
R[upper.tri(R)] ← c(0.5, 0.8, 0.3)
R[lower.tri(R)] ← t(R)[lower.tri(R)]

set.seed(445)

Y ← matrix(0,nrow=n,ncol=k)
X1 ← rbinom(n,1,0.3)
X2 ← rnorm(n,mean =0,sd =1)
```

```
A  ← cbind(1,X1,X2)
e ← rmvt(n,sigma =R,df=7.3)
B ← cbind(beta1,beta2,beta3)
Z ←  A%*% B + log(pt(e,df=7.3)/(1-pt(e,df=7.3)))
Y[Z>0] ← 1

iter ←1050000
nu ← 7.3
betainit ← rep(0,k*c)
rhoinit ← rep(1,n)
Rinit ← rep(0,k*(k-1)/2)
mubeta ← betainit
sigbeta ← c(1000, rep(4, c-1))*diag(k*c)
sigprop ←0.000256*diag(k*(k-1)/2)

as ←Gibbs(A,Y,iter, betainit,rhoinit,Rinit,
    mubeta,sigbeta,sigprop, nu,n,k,c)
```

```
# Lecture des donnees et appelle Gibbs

doses = read.table("doses_reelles.txt",
        sep="\t",dec=",", header=TRUE)

colnames(doses) ← c("ID","Cortico","MTX_iv","MTX_it",
"aracytabine","hydrocor_it","asparaginase","mercapto",
"vincristine","doxorubicin","dexrazoxane")

outcomes ← read.table ("outcomes_covariables.txt",
           sep="\t",dec=",", header=TRUE)

colnames(outcomes) ← c("ID","OBESITY","IR","HTA",
"DYSLIPIDEMIA","HDL_LOW","Age","Age_inter","Temps",
"Sexe","Radiotherapie")

cell = read.table("cellules_blanches.txt",
       sep="\t",dec=".", header=TRUE)

colnames(cell) ← c("ID","WBC_diag")

m ← cbind(outcomes$OBESITY,outcomes$IR,outcomes$HTA,
outcomes$DYSLIPIDEMIA,doses$Cortico, outcomes$Age,
outcomes$Temps, outcomes$Sexe,cell$WBC_diag,
outcomes$Radiotherapie)
```

```
m2 ← na.omit(m)
x ← median(m2[,5])
v← m2[,10] == 0 & m2[,5]> x
m2 ← m2[!v,]

Obesite ←          m2[,1]
IR ←               m2[,2]
HTA ←              m2[,3]
Dyslipidemie ←     m2[,4]
Cortico ←          as.integer(m2[,5]> x)
Age ←              m2[,6]
Temps ←            m2[,7]
Sexe ←             as.integer(m2[,8])
Ncellules ←        m2[,9]
Radiotherapie ← as.integer(m2[,10])

RCor1 ← Cortico ==0 & Radiotherapie ==1
RCor2 ← Cortico ==1 & Radiotherapie ==1
RCor1 ← as.integer(RCor1)
RCor2 ← as.integer(RCor2)

A ← cbind(1,RCor1,RCor2,Ncellules,Age, Temps, Sexe)
Y ← cbind(Obesite, IR, HTA, Dyslipidemie)

iter ←1050000
n ←nrow(A)
k ←ncol(Y)
```

```
c ← ncol(A)
nu ← 7.3

betainit ← rep(0, k*c)
rhoinit ← rep(1,n)
Rinit ← rep(0,k*(k-1)/2)
mubeta ← rep(0, k*c)
sigbeta ← c(1000, rep(4, c-1))*diag(k*c)
sigprop ←0.08* covariance

as←Gibbs(A,Y,iter , betainit ,rhoinit ,Rinit ,
  mubeta,sigbeta ,sigprop , nu,n,k,c)
```

```
# Calcul des differences de risque individuel associees
# au niveau de traitement DF/RT pour chaque facteur de
# risque cardiometabolique.

v ← 1:1000000
sv ← v%%100
sv ← as.logical(sv)
sv← !(sv)
asc ← as[[1]]
asc ← asc[50002:1050001,]
asc ← asc[sv,]

B1 ← A
B1[,2] ←1
```

```
B1[,3] ←0
B2 ← A
B2[,2] ←0
B2[,3] ←0

ATE ← matrix(NA, nrow(A)*10000,4)

for( i in 1:nrow(A)){

  ate1 ← exp(asc[,1:7]%*% B1[i,])/ (1 + exp(asc[,1:7]%*%
    B1[i,]) )

  ate2 ← exp(asc[,1:7]%*% B2[i,])/ (1 + exp(asc[,1:7]%*%
    B2[i,]))

  ate0 ← ate1 -ate2

  ATE [(1+(i-1)*10000): (i*10000),1 ] ← ate0

  ate1 ← exp(asc[,8:14]%*% B1[i,])/ (1 +
    exp(asc[,8:14]%*% B1[i,]) )

  ate2 ← exp(asc[,8:14]%*% B2[i,])/ (1 +
    exp(asc[,8:14]%*% B2[i,]))

  ate0 ← ate1 -ate2

  ATE [(1+(i-1)*10000): (i*10000),2 ] ← ate0
```

```
ate1 ← exp(asc[,15:21]%*% B1[i,])/ (1 +
   exp(asc[,15:21]%*% B1[i,]) )

ate2 ← exp(asc[,15:21]%*% B2[i,])/ (1 +
   exp(asc[,15:21]%*% B2[i,]))

ate0 ← ate1 -ate2

ATE [(1+(i-1)*10000): (i*10000),3 ] ← ate0

ate1 ← exp(asc[,22:28]%*% B1[i,])/ (1 +
   exp(asc[,22:28]%*% B1[i,]) )

ate2 ← exp(asc[,22:28]%*% B2[i,])/ (1 +
   exp(asc[,22:28]%*% B2[i,]))

ate0 ← ate1 -ate2

ATE [(1+(i-1)*10000): (i*10000),4 ] ← ate0

}
```

```
# Calcul des differences de risque multiple associees
# au niveau de traitement DF/RT

v<- 1:1000000
sv <- v%%100
sv <- as.logical(sv)
sv<- !(sv)
asc1 <- as[[1]]
asc1 <- asc1[50002:1050001,]
asc1 <- asc1[sv,]
asc2 <- as[[2]]
asc2 <- asc2[50002:1050001,]
asc2 <- asc2[sv,]

k <- 4
c <- ncol(A)

ATE <- matrix(NA, nrow(A)*10000,16)
Mat.R <- matrix(0,k,k)
l <-1

for( i in 1:nrow(A)){

  for ( j in 1:10000){
```

```
Mat.R[upper.tri(Mat.R)] ← asc2[j,]
Mat.R[lower.tri(Mat.R)] ← t(Mat.R)[lower.tri(Mat.R)]
diag(Mat.R) ← 1
Mat.R ← 2.389*Mat.R

mu1 ←as.vector(B1[i,]%*%matrix(asc1[j,], nrow = c,
    ncol =k))
mu2 ←as.vector(B2[i,]%*%matrix(asc1[j,], nrow = c,
    ncol =k))

ATE [1,1]← pmvt(lower =rep(0,4), upper =rep(Inf,4),
    delta = mu1, df =7, sigma= Mat.R, type = "shifted")
    -
              pmvt(lower =rep(0,4), upper =rep(Inf,4),
                 delta = mu2, df =7, sigma= Mat.R, type
                 = "shifted")

ATE [1,2]← pmvt(lower =c(0,0,0,-Inf), upper
    =c(Inf,Inf,Inf,0), delta = mu1, df =7, sigma=
    Mat.R, type = "shifted")-
              pmvt(lower =c(0,0,0,-Inf), upper
                 =c(Inf,Inf,Inf,0), delta = mu2, df =7,
                 sigma= Mat.R, type = "shifted")

ATE [1,3]← pmvt(lower =c(0,0,-Inf,0), upper
    =c(Inf,Inf,0,Inf), delta = mu1, df =7, sigma=
    Mat.R, type = "shifted")-
              pmvt(lower =c(0,0,-Inf,0), upper
```

```
                =c(Inf,Inf,0,Inf), delta = mu2, df =7,
                sigma= Mat.R, type = "shifted")

ATE [1,4]<- pmvt(lower =c(0,-Inf,0,0), upper
   =c(Inf,0,Inf,Inf), delta = mu1, df =7, sigma=
   Mat.R, type = "shifted")-
            pmvt(lower =c(0,-Inf,0,0), upper
               =c(Inf,0,Inf,Inf), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,5]<- pmvt(lower =c(-Inf,0,0,0), upper
   =c(0,Inf,Inf,Inf), delta = mu1, df =7, sigma=
   Mat.R, type = "shifted")-
            pmvt(lower =c(-Inf,0,0,0), upper
               =c(0,Inf,Inf,Inf), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,6]<- pmvt(lower =c(0,0,-Inf,-Inf), upper
   =c(Inf,Inf,0,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(0,0,-Inf,-Inf), upper
               =c(Inf,Inf,0,0), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,7]<- pmvt(lower =c(0,-Inf,0,-Inf), upper
   =c(Inf,0,Inf,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(0,-Inf,0,-Inf), upper
```

```
                =c(Inf,0,Inf,0), delta = mu2, df =7,
                sigma= Mat.R, type = "shifted")

ATE [1,8]← pmvt(lower =c(0,-Inf,-Inf,0), upper
   =c(Inf,0,0,Inf), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(0,-Inf,-Inf,0), upper
               =c(Inf,0,0,Inf), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,9]← pmvt(lower =c(-Inf,0,0,-Inf), upper
   =c(0,Inf,Inf,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(-Inf,0,0,-Inf), upper
               =c(0,Inf,Inf,0), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,10]← pmvt(lower =c(-Inf,0,-Inf,0), upper
   =c(0,Inf,0,Inf), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
             pmvt(lower =c(-Inf,0,-Inf,0), upper
                =c(0,Inf,0,Inf), delta = mu2, df =7,
                sigma= Mat.R, type = "shifted")

ATE [1,11]← pmvt(lower =c(-Inf,-Inf,0,0), upper
   =c(0,0,Inf,Inf), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
             pmvt(lower =c(-Inf,-Inf,0,0), upper
```

```
                =c(0,0,Inf,Inf), delta = mu2, df =7,
                sigma= Mat.R, type = "shifted")

ATE [1,12]← pmvt(lower =c(0,-Inf,-Inf,-Inf), upper
   =c(Inf,0,0,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(0,-Inf,-Inf,-Inf), upper
               =c(Inf,0,0,0), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,13]← pmvt(lower =c(-Inf,0,-Inf,-Inf), upper
   =c(0,Inf,0,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(-Inf,0,-Inf,-Inf), upper
               =c(0,Inf,0,0), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,14]← pmvt(lower =c(-Inf,-Inf,0,-Inf), upper
   =c(0,0,Inf,0), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(-Inf,-Inf,0,-Inf), upper
               =c(0,0,Inf,0), delta = mu2, df =7,
               sigma= Mat.R, type = "shifted")

ATE [1,15]← pmvt(lower =c(-Inf,-Inf,-Inf,0), upper
   =c(0,0,0,Inf), delta = mu1, df =7, sigma= Mat.R,
   type = "shifted")-
            pmvt(lower =c(-Inf,-Inf,-Inf,0), upper
```

```
                    =c(0,0,0,Inf), delta = mu2, df =7,
                    sigma= Mat.R, type = "shifted")

    ATE [l,16]<- pmvt(lower =c(-Inf,-Inf,-Inf,-Inf), upper
      =c(0,0,0,0), delta = mu1, df =7, sigma= Mat.R, type
      = "shifted")-
                pmvt(lower =c(-Inf,-Inf,-Inf,-Inf),
                    upper =c(0,0,0,0), delta = mu2, df =7,
                    sigma= Mat.R, type = "shifted")
    l <- l+1

  }
}
x <- 1:9900
xb <- rep(x,nrow(A))
ATEb<- cbind(xb,ATE)
ATEb <- data.frame(ATEb)
ATEc<- aggregate(ATEb[,-1], by =list(xb = ATEb[,1]),FUN
  =mean)
ATEc <- ATEc[,-1]

ATEbc <- matrix(0, nrow(ATEc),4)
ATEbc[,1] <- -ATEc[,16]
ATEbc[,2] <-
  -(ATEc[,16]+ATEc[,15]+ATEc[,14]+ATEc[,13]+ATEc[,12])
ATEbc[,3] <- ATEc[,1]+ATEc[,2]+ATEc[,3]+ATEc[,4]+ATEc[,5]
ATEbc[,4] <- ATEc[,1]
```

Algorithme M.C.M.C.

```
// [[Rcpp::export]]
double dmvnrm_arma(rowvec x, rowvec mean,  mat sigma,
                        bool logd = false) {

  int xdim = x.n_elem;
  double out;
  mat rooti = inv(trimatu(chol(sigma)));
  rowvec z    = (x-mean)*rooti;

  out  =  -(static_cast<double>(xdim)/2.0) *
     log(2.0*datum::pi) - 0.5 *sum(z%z) +
     sum(log(rooti.diag()));

  if (logd == false) {
    out = exp(out);
  }
  return(out);
}

// [[Rcpp::export]]
List Gibbs(mat X, mat Y,int iter , vec betainit ,vec
   rhoinit ,vec Rinit ,vec mubeta,
          mat sigbeta ,mat sigprop , double nu,int n,int k,
             int c) {
```

```
//Declaration et initialization de variables

double sig = pow(datum::pi,2.0)*(nu-2)/(3*nu);
double f   = (nu + k)/2;
int Racc = 0;

mat Z(n,k, fill::zeros);
mat za(1,k, fill::zeros);
mat beta(iter+1, k*c, fill::zeros);
mat rho(n,1, fill::zeros);
mat R(iter+1,k*(k-1)/2, fill::zeros);
mat MatR(k,k, fill::zeros);
rowvec mu(k);
rowvec a(k, fill::zeros);
rowvec b(k, fill::zeros);
double s = 0;
double sa=0;
mat W(k*c,k*c, fill::zeros);
mat V(k*c,1, fill::zeros);
mat sigbetatilde(k*c,k*c, fill::zeros);
vec mubetatilde(k*c, fill::zeros);
mat MatRprop(k,k, fill::zeros);
vec Rprop(k*(k-1)/2, fill::zeros);
double num =0;
double dem=0;
mat Rchol(k,k, fill::zeros);
```

```
mat AUX1(k,k, fill::zeros);
mat AUX2(k,k,  fill::zeros);
vec indices=linspace<vec>(1,k,k);

AUX1.each_col() += indices;
AUX2.each_row() += trans(indices);

umat AUX3 = AUX2 > AUX1;

uvec aux= find(AUX3);

beta.row(0) = betainit.t();
rho.col(0)  = rhoinit;
R.row(0)    = Rinit.t();

MatR.elem(aux) =trans(R.row(0));
MatR = symmatu(MatR);
MatR.diag().ones();

mat sigbetainv(inv_sympd(sigbeta));
mat MatRinv(k,k, fill::zeros);
mat MatRinvb(k,k, fill::zeros);

MatRinv = inv_sympd(MatR);

double mu_h = 0.0;
double ab = 0.0;
double ld =0.0;
```

```
double x =0.0;
double u =0.0;

vec sd(k);
mat P(k,k-1,fill::zeros);

umat  AUXb(k-1,k, fill::zeros);

uword uk = static_cast<uword>(k);

AUXb(0,0) = 1;

for ( uword i =1;  i < uk -1; ++i){

  AUXb.row (i) = AUXb.row (i-1) +1;
  AUXb(i,i)  = AUXb(i,i) + 1;
}

umat AUX2b(1,k,fill::zeros);
urowvec au = linspace<urowvec>(0, k-1, k);
AUX2b.row(0)= au;

umat AUX3b(1,n,fill::zeros);
 urowvec av = linspace<urowvec>(0, n-1, n);
AUX3b.row(0)= av;

uword uiter = static_cast<uword>(iter);
```

```
uword un = static_cast<uword>(n);

//Boucle principal

for(uword t=1; t <= uiter; ++t ){

  //Mise a jour pour les zetas (etape Gibbs)

  for(uword i=0; i < un; ++i){

   mu = X.row(i)*reshape(beta.row(t-1),c,k);
   MatRinvb = rho(i,0)/sig*MatRinv;

   a.zeros();
   a(find(Y.row(i)==a)).fill(-datum::inf);

    for(uword h=0; h < uk ; ++h){

     sd(h)     =  sqrt(1/MatRinvb(h,h));
    P.row(h) =  MatRinvb.submat(AUX2b.col(h),
       AUXb.col(h))/MatRinvb(h,h);

   }

   for( uword h=0; h < uk ; ++h){


    mu_h = mu(h) - as_scalar(P.row(h)*
```

```
(trans(Z.submat(AUX3b.col(i),AUXb.col(h))) -
  mu(AUXb.col(h))));

if(a(h) == - datum::inf){

  sa =-1;

}else sa=1;

ab= -sa*mu_h/sd(h);

if(ab < 0) {

  x =rnorm(1,0.0,1.0)[0];

  while( x <ab){

    x =rnorm(1,0.0,1.0)[0];

  }

}else if(ab >=0 && ab < 0.2570){

  x =rnorm(1,0.0,1.0)[0];

  while( fabs(x) <ab){

    x =rnorm(1,0.0,1.0)[0];
```

```
        }

        x = fabs(x);

      }else {

        ld = (ab + sqrt(pow(ab,2.0)+4))/2;

        x =ab + rexp(1,ld)[0];
        u = runif(1,0.0,1.0)[0];

        while( u > exp(-pow(x-ld,2)/2)){

          x =ab + rexp(1,ld)[0];
          u = runif(1,0.0,1.0)[0];

        }
      }

     Z(i,h) = sd(h)*sa*x +mu_h;
    }

  }

  // Mise a jour pour les rhos (etape Gibbs)

  for(uword i=0; i < un; ++i){
```

```
    za = Z.row(i) - X.row(i)*reshape(beta.row(t-1),c,k);
    s =(nu + 1/sig*as_scalar(za*MatRinv*trans(za)))/2;
    rho(i,0) = rgamma(1,f,1/s)[0] ;

  }

// Mise a jour pour les betas (etape Gibbs)

W.zeros();
V.zeros();

for(uword i=0; i < un; ++i){

  W = W + rho(i,0)*
   kron(MatRinv,trans(X.row(i))*X.row(i));

  V = V + rho(i,0)*
   vectorise(trans(X.row(i))*Z.row(i)*MatRinv);

}

sigbetatilde = inv_sympd(sigbetainv + 1/sig*W);
mubetatilde = sigbetatilde*
            (sigbetainv*mubeta +1/sig*V );

beta.row(t)=  mubetatilde.t() +
   randn(1,sigbetatilde.n_cols)*arma::chol(sigbetatilde);
```

```
// Mise a jour pour R (etape Metropolis avec lois de
   proposition uniforme)

 Rprop = trans(R.row(t-1) +
    randn(1,sigprop.n_cols)*arma::chol(sigprop));
 MatRprop.elem(aux) = Rprop;
MatRprop = symmatu(MatRprop);
MatRprop.diag().ones();

if(arma::chol(Rchol,MatRprop) &&
   all(vectorise(abs(MatRprop) <= 1))){

  num = 0;
  dem = 0;

  for(uword i=0; i < un; ++i){

    num = num + dmvnrm_arma(Z.row(i),X.row(i)*
    reshape(beta.row(t),c,k),sig/rho(i,0)*MatRprop,true);

    dem = dem + dmvnrm_arma(Z.row(i),X.row(i)*
    reshape(beta.row(t),c,k),sig/ rho(i,0)*MatR,true);

  }

  if(runif(1)[0] < exp(num - dem)){
```

```
            R.row(t)= trans(Rprop);
            MatR = MatRprop;
           MatRinv = inv_sympd(MatR);
           Racc = Racc +1;

       } else R.row(t) = R.row(t-1);

    }   else R.row(t) = R.row(t-1);

  }
  List z = List::create(beta, R, Racc);
  return z;
}
```

RÉFÉRENCES

Albert, J. et Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.*, *88*(422), 669–79.

Allard, P. *et al.* (2003). Distribution of fasting plasma insulin, free fatty acids, and glucose concentrations and of homeostasis model assessment of insulin resistance in a representative sample of Quebec children and adolescents. *Clin Chem*, *49*(4), 644–649.

Bulow, J., Link, K., Ahrén, B., Nilsson, A-S. et Erfurth, E.-M. (2004). Survivors of childhood acute lymphoblastic leukaemia, with radiation-induced GH deficiency, exhibit hyperleptinaemia and impaired insulin sensitivity, unaffected by 12 months of GH treatment. *Clinical endocrinology*, *61*(6), 683–691.

Castillo, E., Sarabia, J.-M., et Hadi, A.-S. (1997). Fitting continuous bivariate distributions to data. *The Statistician*, *46*(3), 355–369.

Chen, J.,Wildman, R.-P., Hamm, L.-L., Muntner, P., Reynolds, K., Whelton, P.-K. et He, J. (2004). Association between inflammation and insulin resistance in U.S. nondiabetic adults : results from the Third National Health and Nutrition Examination Survey. *Diabetes care*, *27*(12), 2960–2965.

Chib, S. et Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

Gelfand, A.-E. et Smith, A.-F.-M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, *85*(410), 398–409.

Geman, S. et Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans.Pattn. Anal. Mach. Intel.*, *6*(6), 721–741.

Genest, J. *et al.* (2009). Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult - 2009 recommendations. *The Canadian journal of cardiology*, *25*(10), 567–579.

Geweke, J. (1991, avril). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of

constraint probabilities. Dans E. M. Kelamidas (dir.). *Computing Science and Statistics : Proceedings of the 23rd Symposium on the Interface* (p. 571–578). Fairfax Station, VA : Interface Foundation of North America, Inc.

Gumbel, E.-J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association*, *56*, 335–349.

Haddy, T.-B., Mosher, R.-B., Reaman, G.-H. (2009). Late effects in long term survivors after treatment for childhood acute leukemia. *Clinical Pediatrics*, *48*(6), 601–608.

Hardy, O.-T., Czech, M.-P. et Corvera, S. (2012). What causes the insulin resistance underlying obesity? *Current Opinion in Endocrinology, Diabetes, and Obesity*, *19*(2), 81–87. http ://doi.org/10.1097/MED.0b013e3283514e13

Hastings, W.-K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Hund, L. *et al.* (2015). A Bayesian framework for estimating disease risk due to exposure to uranium mine and mill waste on the Navajo Nation. *Journal of the Royal Statistical Society : Series A (Statistics in Society),178*(4),1069–1091. `http://dx.doi.org/10.1111/rssa.12099`

Kahn, B.-B. et Flier, J.-S. (2000). Obesity and insulin resistance. *Journal of Clinical Investigation*, *106*(4), 473–481. `http://doi.org/10.1097/MED.0b013e3283514e13`

Katzmarzyk, P.-T. (2004). Waist circumference percentiles for Canadian youth 11-18y of age. *Eur J Clin Nutr*, *58*(7), 1011–1015, `http://doi:10.1038/sj.ejcn.1601924`

Klop, B., Elte, J.-W.-F., et Castro Cabezas, M. (2013). Dyslipidemia in Obesity : Mechanisms and Potential Targets. *Nutrients*, *5*(4), 1218–1240. `http://doi.org/10.3390/nu5041218`

Li, Y., Ghosh, S. K. (2013). *Efficient sampling method for truncated multivariate normal and Student t-distribution subject to linear inequality constraints : rapport technique.* North Carolina State University. Recupéré de `http://www.stat.ncsu.edu/information/library/papers/mimeo2649_Li.pdf`

Lughetti, L., Bruzzi, P., Predieri, B., Paolucci, P. (2012). Obesity in patients with acute leukemia in childhood. *Ital J Pediatr.*, *38*(4). `http://doi:10.1186/1824-7288-38-4`

Malik, H.-J. et Abraham, B.(1973). Multivariate logistic distributions. *The Annals of Statistics 1*(3), 588–590.

Marcoux, S. *et al.* (2016). The PETALE study : Late adverse effects and biomarkers in childhood acute lymphoblastic leukemia survivors. *Pediatr Blood Cancer* ,*64*(6), http://doi:10.1002/pbc.26361

Metropolis, N., Rosenbluth, A.-W., Rosenbluth, M.-N., Teller, A.-H. et Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087–1092.

Mody, R. *et al.* (2008). Twenty-five-year follow-up among survivors of childhood acute lymphoblastic leukemia : a report from the Childhood Cancer Survivor Study. *Blood*, *111*(12), 5515–5523.

Natarajan, R. et Kass, R.-E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*(449), 227-237.

Natarajan, R., McCulloch, C.-E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, *82*(3), 639-643. https://doi.org/10.1093/biomet/82.3.639

Nottage, K.-A., Ness, K.-K., Li, C., Srivastava, D., Robison, L.-L. et Hudson, M.-M. (2014). Metabolic Syndrome and Cardiovascular Risk among Long-Term Survivors of Acute Lymphoblastic Leukaemia - From the St. Jude Lifetime Cohort. *British Journal of Haematology*, *165*(3), 364–374. http://doi.org/10.1111/bjh.12754

Nummelin, E. (1984). *General Irreductible Markov Chains and Non-Negative Operators*. Cambridge : Cambridge Univ. Press.

O'Brien, S.-M. et Dunson, D.-B. (2004). Bayesian multivariate logistic regression. *Biometrics*, *60*(3), 739–746.

Olefsky J. M., Glass C. K. (2010). Macrophages, inflammation, and insulin resistance. *Annual Review of Physiology*, *72*, 219-46. https://doi.org/10.1146/annurev-physiol-021909-135846

Padhye B., Dalla-Pozza, L., Little, D., Munns, C. (2016). Incidence and outcome of osteonecrosis in children and adolescents after intensive therapy for acute lymphoblastic leukemia (ALL). *Cancer Medicine*, *5*(5), 960-967. http://doi.org/10.1002/cam4.645

Paradis, G., Tremblay, M.-S., Janssen, I., Chiolero, A. et Bushnik, T. (2010). Blood pressure in Canadian children and adolescents. *Health reports*, *21*(2), 15–22.

Pui, C. H. *et al.* (2012). Pediatric acute lymphoblastic leukemia : where are we going and how do we get there?. *Blood*, *120*(6), 1165–1174.

Revuz, D. (1975). *Markov Chains*. Amsterdam : North-Holland.

Robert, C.-P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, *5*(2), 121–125.

Roberts, G.-O. et Smith, A.-F.-M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, *49*(2), 207–216. `https://doi.org/10.1016/0304-4149(94)90134-1`

Rosenthal, J.(2011). Optimal Proposal Distributions and Adaptative MCMC. [Chapitre de livre]. Dans S. Brooks *et al.* (dir.), *Handbook of Markov Chain Montecarlo* (p. 93-110). Hoboken : Chapman & Hall/CRC.

Rousseeuw, P. et Molenberghs, G. (1994). The shape of correlation matrices. *Am. Statistician*, *48*(4), 276–9.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Stern, R.-H.-T. (2013). The new hypertension guidelines. *J Clin Hypertens (Greenwich)*, *15*(10), 748–751. `https://doi.org/10.1111/jch.12171`

Stiralli, R., Laird, N. et Ware, J. (1984). Random effects models for serial observations with binary responses. *Biometrics*, *40*(4), 961–970.

Tabachnick, B.-G. et Fidell, L.-S. (2013). *Using Multivariate Statistics* (6e éd.). Boston : Pearson.

Tierney, L. (1994). Markov Chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1701– 1762.

Van den Broeck, J., Willie, D. et Younger, N. (2009). The World Health Organization child growth standards : expected implications for clinical and epidemiological research. *European journal of pediatrics*, *168*(2), 247–251, `https://doi.org/10.1007/s00431-008-0796-9`

Vozarova B., Weyer C., Lindsay R.-S., Pratley R.-E., Bogardus C. et Tataranni P.-A. (2002). High white blood cell count is associated with a worsening of insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes*, *51*(2), 455-61. `https://doi.org/10.2337/diabetes.51.2.455`

Waber D. P. *et al.* (2000). Cognitive sequelae in children treated for acute lymphoblastic leukemia with dexamethasone or prednisone. *J. Pediatr. Hematol.*

Oncol., *2000*, *22*3.206–213.

Yang, Q. et Wang, Y. (2012). Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *Journal of Probability and Statistics*, *2012*(ID : 652569). https://doi.org/10.1155/2012/652569

Yoshimura, A. *et al.* (2015). Association of peripheral total and differential leukocyte counts with obesity-related complications in young adults. *Obesity Facts*, *8*(1), 1-16. https://doi.org/10.1159/000373881

Zeger, S.-L. et Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121–130.