Nouveau cadre statistique pour la cartographie-fine

Mémoire

El Hadji Ousseynou Accrachi

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

Québec, Canada

Résumé

Des études d'association à l'échelle du génome (GWAS) ont permis l'identification de milliers de régions du génome comportant des variants génétiques associés à des traits et qui peuvent être à l'origine de certaines maladies complexes. Cependant faire des tests biologiques pour tous les variants génétiques découverts à l'aide de GWAS est pratiquement impossible. Ainsi, les études de cartographie-fine visent à déterminer un ensemble cible de variants génétiques susceptibles d'être associés à un trait d'intérêt. Les principales difficultés pour les méthodes statistiques pour la cartographie-fine sont la présence de milliers de variants génétiques pour seulement une centaine d'individus et la présence d'une forte structure de corrélation, ou déséquilibre de liaison (LD) entre les variants génétiques. Il existe de nombreuses contributions dans les études de cartographie-fine notamment CAVIAR [19], CAVIAR-Gene [30], PAINTOR [28], fastPAINTOR [27] etc. Ces études se basent sur des méthodes statistiques de sélection d'un ensemble crédible de variants génétiques pour aider à prioriser les variants et à discerner les conséquences fonctionnelles du risque de maladies des variants sélectionnés. Dans ce mémoire, nous proposons un nouveau cadre statistique avec une procédure de sélection de variants génétiques (SNPs). Nous utilisons une méthode conditionnelle ou bayésienne pour identifier les SNPs susceptibles d'être causaux. Ainsi la statistique d'association d'un SNP est réécrite et sa loi asymptotique est déterminée. Notre procédure de sélection est itérative et grâce à une loi *a priori*, elle calcule les probabilités *a posteriori* pour qu'un SNP soit significatif pour le trait d'intérêt. À chaque étape les statistiques d'association des SNPs sont calculées et le SNP avec la plus forte probabilité *a posteriori* est choisi. Dans nos simulations, nous montrons que la correction sur la loi asymptotique de la statistique d'association apporte une amélioration significative dans la sélection des SNPs qui ont un lien avec le trait d'intérêt.

Table des matières

A ma brave mère Fatou Ndiaye et
mon père feu Docteur Sourakhé
Cissokho Accrachi

Table de notation

i	Individu
j	SNP
k	Étape d'itération
K	Le nombre de SNP à sélectionner ($k \leq K$)
N	Le nombre d'individus
m	Le nombre de SNPs
G_j	Le code du SNP j, $G_j \in \{0,1,2\}$
X_j	La variable aléatoire centrée et réduite de G_j
Y_j	Phénotype ou trait
Z_j	Statistique d'association du SNP j
$\hat{Z}_{jk}$	Estiamteur de la statistique d'association du SNP j à la k-ème étape

Remerciements

Je remercie d'abord ALLAH Le Tout Puissant et Le Prophète Mohamet (PSL) ainsi que Son Humble Serviteur Serigne Touba Mbacké (RTA).
J'adresse ma gratitude et mes sincères remerciements à mes professeurs M'Hamed Lajmi Lakhal Chaieb et Anne-Sophie Charest qui m'ont donné l'opportinuité de découvrir les vastes champs de recherches en biostatistique. Ils m'ont aidé dans la compréhension de mon sujet de mémoire, guidé dans la méthodologie, dans la recherche, transmis la rigueur dans le travail et soutenu financièrement. Grâce à leur disponibilité et leur encadrement rigoureux dans mes démarches pédagogiques et scientifiques, je suis passé avec succès de la statistique des processus de Poisson à la statistique génétique.
Mes remerciements vont aussi à l'encontre du département des Mathématiques et statistique de l'Université Laval pour la qualité des cours dispensés, pour le cadre idéal d'études et de recherches et pour les tâches d'auxiliaires d'enseignement qui m'ont permis d'acquérir une expérience conséquente en milieu académique.
Je tiens à remercier infiniment mes très chers parents, ma famille, mes amis, mes collègues et mes voisins pour le soutien et les conseils durant toute la période de la préparation de ce mémoire.
Enfin, je termine par remercier les examinateurs de ce document pour le temps précieux dévoué à l'amélioration de ce mémoire.

Introduction

Une étude d'association à l'échelle du génome (GWAS selon son acronyme en anglais) [5], permet de détecter les variants génétiques associés à certains traits. Avec le développement des technologies de séquençage, chaque SNP est génotypé pour savoir s'il est causal dans une maladie [5], [9]. À partir d'un ensemble de données GWAS, les études de cartographie-fine permettent d'identifier, ou de prédire, les SNPs provoquant une maladie. L'approche classique d'une cartographie-fine exécute une procédure de sélection itérative pour prédire les SNPs causaux dans une région du génome. D'abord, la statistique d'association de chaque SNP est calculée. Ensuite le SNP le plus fortement associé est choisi comme SNP causal. Si la région génomique contient un seul SNP causal alors le SNP le plus significativement associé est probablement le SNP causal lui-même [13]. Cependant la région génomique peut contenir plusieurs SNPs causaux et de plus ces SNPs peuvent être corrélés ou en déséquilibre de liaison (LD) [19], [20]. Pour tenir compte du fait que certains SNPs causaux peuvent être en LD avec d'autres SNPs non causaux, l'approche conditionnelle ou Bayésienne d'une cartographie-fine recalcule la statistique d'association des SNPs à chaque itération en tenant compte des SNPs sélectionnés dans les itérations précédentes.

Les méthodes statistiques pour la cartographie-fine ont eu plusieurs succès dans la détection de variants responsables de maladies [10], [13], [19], [30], [27], [28]. Le point commun des études de cartographie-fine est l'utilisation d'un cadre statistique pour la sélection de variants. Dans notre travail, nous nous intéressons d'abord aux calculs des statistiques d'association et de leurs distributions dans les méthodes statistiques de cartographie-fine existantes [19], [30], [27], [28]. Ensuite, nous proposons une procédure conditionnelle permettant de sélectionner un ensemble cible Ψ de SNPs à partir de données sur un trait, de SNPs et de connaissances *a priori* sur ces SNPs.

Plusieurs méthodes de cartographie-fine existantes supposent une distribution multivariée du vecteur des statistiques d'association en considérant que la distribution marginale des statistiques d'association de chaque SNP a asymptotiquement une distribution normale de variance 1 [30], [19], [28], [27]. Nous pouvons toutefois montrer que cette variance asymptotique n'est pas égale à 1. Ensuite, pour calculer la probabilité *a posteriori*, une probabilité *a priori* discrète est proposée [30], [19], [28], [27]. Le calcul de la probabilité *a posteriori* qu'un SNP soit causal est souvent fastidieux avec cette hypothèse [30], [19], [28], [27], [7], [18]. Nous

utilisons par la suite une distribution *a priori* appropriée pour calculer la probabilité qu'un SNP soit causal.

Notre contribution scientifique dans ce travail est de dériver la distribution correcte de la statistique d'association. De plus, grâce à une nouvelle distribution pour la loi *a priori*, nous parvenons à avoir une expression explicite pour la probabilité *a posteriori*. Il faut noter que la méthode de sélection de SNPs proposée ici est valable pour tout autre loi *a priori* choisie.

Dans ce mémoire, au chapitre 1 nous introduisons quelques concepts génétiques et passons en revue les études GWAS et leurs applications. Ensuite le chapitre 2 nous permettra de définir notre modèle statistique et la procédure de sélection de SNPs proposée. Enfin, au chapitre 3 nous introduisons le modèle de génération de données et donnons les résultats de simulations de quelques scénarios pour illustrer les avantages de notre procédure.

Chapitre 1

Notions génétique et GWAS

1.1 Quelques concepts en génétique

Les définitions qui sont données dans les sections suivantes sont essentiellement basées sur le livre de Elizabeth A. Thompson [29].

1.1.1 ADN

Grégor Mendel (1822-1884) est le père précurseur de la génétique avec ses études sur la variété de petits pois. Il montre l'existence d'une transmission des caractères des parents à leurs descendants. La génétique est une branche de la biologie étudiant l'hérédité, elle cherche à identifier les modalités de leur transmission et les caractères héréditaires.
Nous savons que tout être vivant est composé de cellules. Les noyaux des cellules contiennent l'information génétique dite ADN (acide désoxyribonucléique). L'ensemble de l'ADN présent dans le noyau de chacune des cellules d'un organisme constitue le génome. L'ADN est formée de 4 bases différentes (ou nucléotides) que sont l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C). Les paires $A - T$ et $G - C$ sont complémentaires (c'est-à-dire que l'adénine s'apparie toujours avec la tymine, et la guanine avec la cytosine). Une représentation de l'ADN est donnée par la figure 1.1.
L'ADN d'un individu se compose d'environ 3×10^9 paires de bases. Cet ADN est emballé dans des chromosomes dont chacun a une séquence d'ADN dans une structure à double hélice torsadée. Il y a 46 chromosomes dans le noyau de chaque cellule humaine normale, c'est-à-dire 22 paires d'autosomes et une paire de chromosomes sexuels. Sur les deux chromosomes d'une paire, l'un vient de l'ADN de la mère de l'individu et l'autre de l'ADN du père. Les autosomes contiennent en général la majorité de l'ADN codant les protéines qui à leur tour affectent les caractéristiques des individus. Un fragment d'ADN constitue un gène. L'étude Human Genom project qui a été complétée en 2003 a permis d'identifier les 20 000 à 30 000 gènes humains. Les deux chromosomes d'une paire homologue portent les mêmes gènes à des localisation identiques.

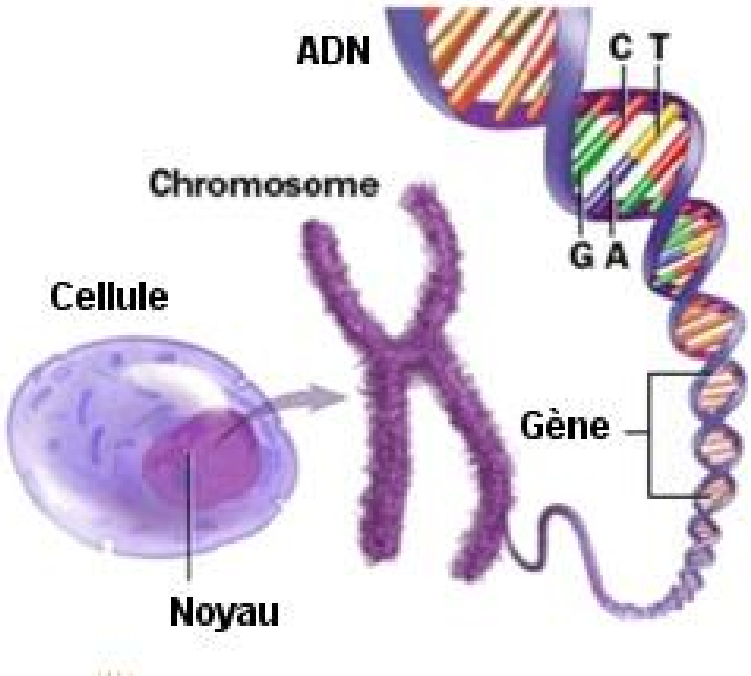

FIGURE 1.1 – Représentation de l'ADN

Source : http ://www.psychomedia.qc.ca/sante/2014-05-08/biologie-synthetique-adnpartiellement- artificiel

Une des différentes versions d'un gène donné est appelée allèle. Un locus présentant une variation non négligeable dans une population est appelé *polymorphisme*.
Un *marqueur génétique* est un polymorphisme qui n'est pas un gène dont on connaît la position sur le génome. Le rôle principal des marqueurs génétiques en statistique génétique est de fournir des points de répérage afin que de nouveaux gènes puissent être localisés. Classiquement, la fréquence du génotype le plus fréquent doit être inférieure à 99% pour qu'un gène soit qualifié de polymorphisme.

1.1.2 SNP

Le SNP (*Single Nucleotide Polymorphism*) est le variant génétique le plus fréquent dans le génome. Il se présente sous la forme de deux allèles qui diffèrent seulement par un seul nucléotide. Par exemple, les deux allèles **...AATGTAAC...** et **...AATGCAAC...** constituent un SNP. La figure 1.2 nous permet d'avoir une visualisation d'un SNP. Le développement des nouvelles technologies de séquençage à haut débit favorise l'accroissement permanent du nombre de variations ou SNPs de la séquence du génome humain [2], [24]. Ainsi les techniques de séquençage permettent aujourd'hui d'identifier environ 4 à 5 millions de SNPs dans le

génome d'une personne [1]. La collection de ces informations est d'un intérêt majeur, non seulement pour la recherche mais également pour le diagnostic des maladies génétiques [2].

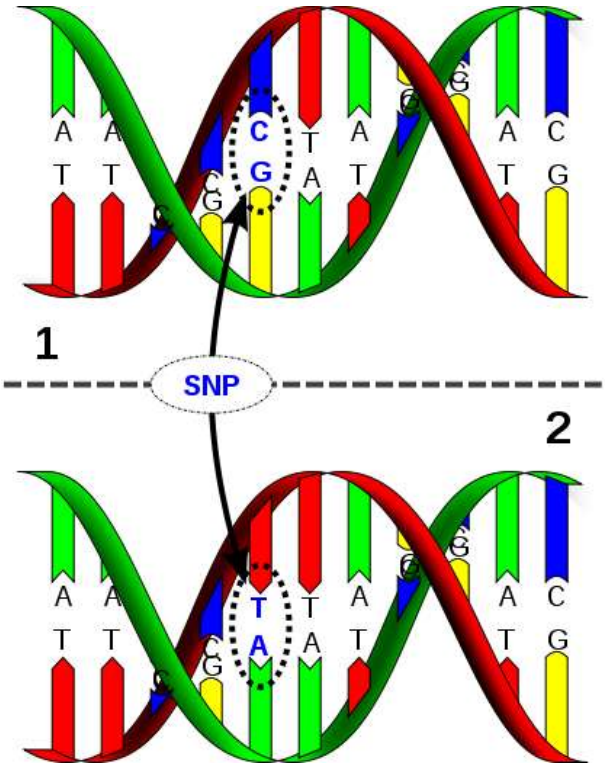

FIGURE 1.2 – Représentation d'un SNP

Source : https ://www.futura-sciences.com/sante/definitions/genetique-snp-6348/

Dans un SNP, l'allèle le moins fréquent dans la population est souvent noté en minuscule et sa fréquence est dite fréquence allélique mineure (MAF). Si la MAF d'un SNP est inférieure à 0.05 (ou inférieure à 0.01 selon la littérature) alors le SNP est dit variant rare, sinon il est dit variant commun.

1.1.3 Génotype et phénotype

Les paires d'allèles non ordonnées qu'un individu possède constituent son génotype. En considérant deux gènes à deux allèles chaque, Aa et Bb, alors il y a neuf génotypes possibles qui sont AA/BB, Aa/BB, aa/BB, AA/Bb, Aa/Bb, aa/Bb, AA/bb, Aa/bb et aa/bb. Si le locus est lié à un gène fonctionnel, la caractéristique potentiellement observable résultante de l'individu est le phénotype. Donc le phénotype ou trait peut être considéré comme l'ensemble des caractéristiques physiques ou physiologiques d'un individu. Il est déterminé par les gènes ou allèles, l'environnement, le mode de vie, etc. Pour les traits génétiques, le phénotype d'un individu est déterminé en partie par son génotype. Dans notre travail, le génotype est le SNP et le phénotype caractérise une maladie complexe.

1.1.4 Fréquence allélique et équilibre de Hardy-Weinberg

La fréquence allélique est la fréquence à laquelle se trouve l'allèle d'un variant dans une population. Considérons un seul locus génétique, avec n allèles A_j et ayant des fréquences alléliques notées par f_j, $j = 1, \cdots, n$. Maintenant, nous supposons que dans une population d'accouplements aléatoires, les types alléliques des gènes maternels et paternels chez un individu sont indépendants. Un individu est homozygote s'il possède deux allèles identiques sur un même gène. Ainsi, la probabilité qu'un individu soit homozygote A_jA_j est f_j^2. Tandis qu'un individu est hétérozygote s'il a deux allèles différentes. La probabilité que l'individu soit hétérozygote A_jA_l, $j < l$ est $2f_jf_l$. Ces fréquences de génotypes sont connues sous le nom de proportions de Hardy-Weinberg, et une population présentant des génotypes dans ces proportions est dite en équilibre de Hardy-Weinberg.

1.1.5 Déséquilibre de liaison

Le déséquilibre de liaison ou LD mesure la corrélation entre deux variants génétiques ou SNPs. Il peut être vu comme une mesure de corrélation entre deux loci. Soit les deux SNPs représentés par leurs allèles Aa et Bb, une mesure de l'association allélique entre ces deux SNPs ou entre les loci A et B est

$$\Delta = \mathbb{P}(AB) - \mathbb{P}(A)\,\mathbb{P}(B)\,. \tag{1.1}$$

Cette mesure est due à Robbins [26] et est connue sous le nom de *coefficient de déséquilibre de liaison*.
Si $\Delta = 0$, nous disons que les deux SNPs ne sont pas corrélés. Cependant, si $|\Delta|$ est grand (respectivement petit) alors la corrélation entre les deux SNPs est forte (respectivement faible).

1.2 GWAS et applications

Une étude d'association à l'échelle du génome (GWAS) est une étude des variants génétiques communs à l'ensemble du génome humain conçue pour identifier les associations génétiques avec des traits observables [31]. Les GWAS ont permis l'identification de milliers de régions du génome avec des variants génétiques susceptibles de maladies [11], [10], [17]. Cependant, seulement une petite partie des variants génétiques a été validée par les biologistes comme étant réellement associées aux maladies génétiques [23]. La recherche de variants associés à un trait est souvent une tâche difficile du fait du nombre important de variants présents dans le génome et de la corrélation qui peut exister entre ces variants [25], [8], [16]. Le déséquilibre de liaison (LD) mesure la corrélation entre des variants. Dans les études de GWAS, la présence de LD entre les SNPs peut causer la sélection de SNPs qui ne sont pas significativement liés au trait d'intérêt mais qui sont en corrélation au vrai SNP causant.
Dans l'optique de capturer les variants causaux, les nouvelles techniques de séquençage à haut débit ont permis le génotypage d'un très grand nombre de variants génétiques. L'association

à un trait des SNPs sera donc testée individuellement [28], [19].
Les études de cartographie-fine fournissent des approches d'identification de gènes associés à des traits d'intérêts. Elles suivent généralement une procédure en deux étapes. Tout d'abord, une analyse statistique du signal d'association est effectuée pour identifier un ensemble minimum de SNPs pouvant expliquer le signal. Deuxièmement, les SNPs qui sont supposés causaux sont testés en utilisant des tests fonctionnels laborieux et coûteux. Donc les méthodes statistiques de cartographie-fine ont pour objectif de minimiser le nombre de SNPs qui doivent être sélectionnés pour les tests fonctionnels [28], [30], [19].
La base des études statistiques en cartographie-fine est la sélection de SNPs pour la validation fonctionnelle grâce à la force du signal d'association. Par un test d'association statistique, les «K meilleurs SNPs» sont sélectionnés pour les tests fonctionnels. Malheureusement, la présence du déséquilibre de liaison peut induire des statistiques d'associations plus élevées pour les SNPs voisins des variants causaux [15]. Ce qui fait que cette approche ne donne aucune garantie que les SNPs causaux réels sont contenus dans les «K meilleurs SNPs» sélectionnés. Les travaux [13], [6] proposent d'estimer la probabilité de chaque SNP d'être causal à un locus donné sous l'hypothèse que le locus ne peut contenir plus d'un variant causal. Ainsi la probabilité *a posteriori* d'être causal est calculée en utilisant seulement les statistiques d'association de chaque SNP. Un inconvénient de cette approche est que les loci à risque peuvent héberger plus d'un variant causal [12], [4], [14]. Pour tenir compte de cette limite, des approches conditionnelles ont été proposées [12], [4], [14]. L'approche conditionnelle sélectionne itérativement la plupart des SNPs associés puis elle recalcule la valeur de la statistique pour les SNPs restants conditionnellement aux SNPs déjà sélectionnés. Les itérations se terminent lorsqu'une condition d'arrêt prédéterminée est satisfaite. Ceci permet d'avoir la probabilité que chaque variant soit causal sachant qu'un locus peut contenir un nombre arbitraire de variants causaux [19]. L'approche conditionnelle peut aussi exclure des variants causaux en fonction de l'ordre dans lequel les SNPs sont sélectionnés dans la procédure d'itération.
Dans le processus d'identification de variants associés à un trait, nous notons quatre catégories de variants : les variants testés significatifs et causaux (variants réellement causaux), ceux testés non significatifs et causaux (variants causaux non détectés), ceux testés significatifs et non causaux (variants faussement causaux) et ceux testés non significatifs et non causaux (variants non causaux).

Chapitre 2

Procédure de sélection de SNPs

La statistique génétique est une branche de la statistique qui consiste à développer des outils d'analyse de données génomiques provenant de diverses plateformes moléculaires [22]. Les méthodes statistiques pour la cartographie-fine ont pour objectifs d'identifier ou de prédire les SNPs provoquant une maladie à partir d'un ensemble de données GWAS donné. Cependant nous pouvons citer quelques facteurs qui font que les méthodes de régression classique ne sont pas souvent adéquates :

- le nombre de SNPs présent dans GWAS est trop grand pour tous les tester en laboratoire,
- le nombre de SNPs dans GWAS est largement plus grand que le nombre d'individus,
- la grande dimension des GWAS amène également un problème de redondance au sein des données,
- l'existance de structures de corrélation ou déséquilibre de liaison entre les SNPs.

La méthode utilisée dans les études de cartographie-fine consiste à faire des régressions individuelles sur chaque SNP. Ensuite, les SNPs sont classés selon leur force d'association à la maladie. Cette méthode a donné des résultats probants pour aider à sélectionner un ensemble de SNPs qui causent certaines maladies. L'inconvénient de cette méthode est que les SNPs les plus fortement associés sont susceptibles d'être en LD avec le «vrai» SNP causal plutôt que d'avoir une fonction biologique significative sur la maladie. Le défi est de trouver une procédure efficiente qui nous permettra de sélectionner ou capturer les «vrais» SNPs causals.

2.1 Données et modèle

Dans ce mémoire, nous utilisons la variante la plus commune du génome appelée polymorphysme mononucléotidique (SNP) pour identifier les variants génétiques associés au trait d'intérêt (maladie complexe). Considérons que nous avons un phénotype ou trait Y sur N individus et m génotypes ou SNPs. Le nombre d'individus est souvent inférieur aux variables explicatives ($N \lll m$). Cependant, seulement un sous-ensemble des m SNPs est significativement associé au trait Y.

Pour un individu i et un SNP j dont la valeur G_{ij} est codé additivement comme 0, 1 ou 2 pour le nombre d'allèles mineurs ($G_{ij} \in \{0,1,2\}$). Sans perte de généralité, nous pouvons normaliser le code des SNPs par $X_{ij} = \left(G_{ij} - \overline{G}_j\right) / \sqrt{S^2_{G_j}}$ avec $\overline{G}_j = N^{-1} \sum_{k=1}^{N} G_{ik}$ et $S^2_{G_j} = N^{-1} \sum_{i=1}^{N} \left(G_{ij} - \overline{G}_j\right)^2$.
Donc $\sum_{i=1}^{N} X_{ij} = 0$ et $\sum_{i=1}^{N} X^2_{ij} = N$ pour $j = 1, \cdots, m$, $X_j = (X_{1j}, \cdots, X_{Nj})^T$ et $Y = (Y_1, \cdots, Y_N)^T$.
L'approche standard pour illustrer la relation entre un SNP et le trait d'intérêt (souvent une maladie complexe) est à travers un modèle de régression simple. En considérant un individu i dont le trait d'intérêt Y_i a été mesuré et $X_i = (X_{i1}, ..., X_{im})^T$ le vecteur de SNPs, alors le trait peut être modélisé par

$$Y_i = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \tag{2.1}$$

où $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2_j\right)$, $\sigma_j > 0$ est l'écart type de l'erreur aléatoire due aux facteurs environnementaux et β est le vecteur des coefficients de la régression. La taille de l'effet du SNP j est mesurée par β_{1j} et son estimateur (dans ce mémoire nous prenons l'estimateur des moindre carrés) est $\hat{\beta}_{1j} = X_i^T Y / N$.
La force d'association du SNP j au trait Y est ensuite quantifiée en utilisant la statistique de Wald [3]

$$\hat{Z}_j = \frac{\hat{\beta}_{1j}}{\widehat{SE}\left(\hat{\beta}_{1j}\right)} \tag{2.2}$$

avec $\widehat{SE}\left(\hat{\beta}_{1j}\right) = \sqrt{\widehat{Var}\left(\hat{\beta}_{1j}\right)}$ où $\widehat{Var}\left(\hat{\beta}_{1j}\right) = \hat{\sigma}^2_j / N$. Dans notre travail, la variance des termes d'erreur est estimée par l'estimateur du maximum de vraisemblance qui est défini comme suit :

$$\hat{\sigma}^2_j = \frac{1}{N-2} \left(Y - X\hat{\beta}_j\right)^T \left(Y - X\hat{\beta}_j\right)$$

avec $\hat{\beta}_j = \left(\hat{\beta}_{0j}, \hat{\beta}_{1j}\right)^T$ et $X = \begin{pmatrix} 1 & X_{1j} \\ \vdots & \vdots \\ 1 & X_{Nj} \end{pmatrix}$.
Pour $j = 1, \cdots, m$, nous introduisons le paramètre λ_j qui mesure la force de dépendance entre le trait Y et le SNP j. Nous supposons que $\lambda_j = 0$ correspond à un SNP j non causant. Par contre $\lambda_j > 0$ (respectivement $\lambda_j < 0$) signifie une dépendance positive (respectivement négative) entre Y et le SNP j.

De plus, nous supposons une loi *a priori* pour λ_j. Plusieurs choix de lois *a priori* pour λ ont été proposés dans la littérature [28]. Dans ce mémoire, nous suivons [28] et nous proposons une loi à deux composantes

$$p_j = \mathbb{P}\left(\lambda_j = 0\right) \tag{2.3}$$

$$\text{et} \quad \lambda_j \mid \lambda_j \neq 0 \sim \mathcal{N}\left(0, \sigma^2_{\lambda_j}\right), \quad \sigma_{\lambda_j} > 0. \tag{2.4}$$

La loi *a priori* 2.4 est utilisée pour la dérivation de la probabilité *a posteriori* que le SNP soit causant. Même si un seul choix de loi *a priori* est considéré dans ce mémoire, notre développement théorique peut facilement s'adapter à n'importe quelle distribution.
Dans ce mémoire, nous proposons une procédure de sélection de SNP par étape. À chaque étape, nous sélectionnons le SNP ayant la plus forte probabilité *a posteriori*

$$q_j = \mathbb{P}\left(\lambda_j \neq 0 \mid \text{modèle, données}\right). \tag{2.5}$$

Nous présentons ci-dessous les détails des étapes de notre procédure de sélection de SNPs.

2.2 Étapes de sélection de SNPs

Dans notre travail, nous considérons que le trait Y est donné ainsi que les m SNPs. Pour chaque étape, nous considérons plusieurs modèles dont le trait Y est indépendant des SNPs.

2.2.1 Étape 1

À l'étape 1, on ajuste un modèle de régression simple pour chacun des m SNPs. Nous considèrons les modèles suivants

$$Y_i = \beta_{0j1} + \beta_{1j1} X_{ij1} + \varepsilon_{ij1}, \qquad j = 1,...,m$$

où $\varepsilon_{ij1} \sim \mathcal{N}\left(0,\sigma^2_{j1}\right)$, $\sigma_{j1} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape 1, $\forall i = 1, \cdots, N$ et $X_{ij1} = X_{ij}$ (nous reviendrons sur cette notation aux étapes suivantes).
La statistique d'association pour mesurer l'effet du SNP j est définie par

$$\hat{Z}_{j1} = \frac{\hat{\beta}_{1j1}}{\sqrt{\hat{\sigma}^2_{j1}/N}} \tag{2.6}$$

avec $\hat{\sigma}^2_{j1}$ est l'estimateur de la variance de l'erreur aléatoire du modèle de la première étape. $\hat{\sigma}^2_{j1}$ est définie comme suit :

$$\hat{\sigma}^2_{1j} = \frac{1}{N-2}\left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)^T \left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)$$

avec $\hat{\beta}_{j1} = \left(\hat{\beta}_{0j1},\hat{\beta}_{1j1}\right)^T$ et $\tilde{X}_{j1} = \begin{pmatrix} 1 & X_{1j} \\ \vdots & \vdots \\ 1 & X_{Nj} \end{pmatrix}$.
Dans l'annexe A.1, nous montrons que

$$\hat{Z}_{j1} \sim \mathcal{N}\left(\sqrt{N}\lambda_{j1}, 1 + \frac{1}{2}\lambda^2_{j1}\right) \tag{2.7}$$

où $\lambda_{j1} = \beta_{1j1}/\sqrt{\sigma^2_{j1}}$.

Dans les travaux antérieurs, [28], [30], [27], et [21] ont ignoré la variabilité de $\hat{\sigma}^2_{ij1}$ et ont simplement considéré $\text{Var}(\hat{\beta}_{1j1}/\sqrt{\sigma^2_{ij1}/N}) = 1$. Dans les étapes suivantes, on verra que la correction de cette variabilité influence aussi les formules de covariances entre des statistiques de tests d'intérêt. La principale contribution méthodologique de ce mémoîre est d'étudier l'impact de considérer la variabilité des termes d'erreurs sur la procédure de sélection de SNPs. En effet, les corrections introduites dans ce mémoire ont un impact sur les probabilités *a posteriori*.

Dans ce mémoire, nous supposons que la loi *a priori* est la même pour toutes les étapes de la procédure de sélection. La loi *a priori* choisie dans ce travail est la suivante :

$$\begin{aligned} p_j &= \mathbb{P}\left(\lambda_{j1} = 0\right) \\ \text{et} \quad \lambda_{j1} \mid \lambda_{j1} \neq 0 &\sim \mathcal{N}\left(0, \sigma^2_{\lambda_j}\right), \quad \sigma_{\lambda_j} > 0. \end{aligned}$$

La probabilité *a posteriori* pour mesurer l'implication du SNP j est donc donnée par

$$\begin{aligned} q_j &= \mathbb{P}\left(\lambda_{j1} \neq 0 \mid \hat{Z}_{j1}\right) \\ &= \frac{\mathbb{P}\left(\lambda_{j1} \neq 0\right) \mathbb{P}\left(\hat{Z}_{j1} \mid \lambda_{j1} \neq 0\right)}{\mathbb{P}\left(\lambda_{j1} \neq 0\right) \mathbb{P}\left(\hat{Z}_{j1} \mid \lambda_{j1} \neq 0\right) + \mathbb{P}\left(\lambda_{j1} = 0\right) \mathbb{P}\left(\hat{Z}_{j1} \mid \lambda_{j1} = 0\right)} \\ &= \frac{\left(1 - p_j\right) A_{1j1}}{\left(1 - p_j\right) A_{1j1} + p_j A_{2j1}} \end{aligned}$$

où $A_{1j1} = \mathbb{P}\left(\hat{Z}_{j1} \mid \lambda_{j1} \neq 0\right)$ et $A_{2j1} = \mathbb{P}\left(\hat{Z}_{j1} \mid \lambda_{j1} = 0\right)$.
Pour ne pas alourdir nos calculs, nous considérons que les réalisations des variables aléatoires $\hat{Z}_{j1}$ et λ_{j1} sont aussi notées respectivement par $\hat{Z}_{j1}$ et λ_{j1}. Nous notons la densité de $\hat{Z}_{j1} \mid \lambda_{j1} \neq 0$ par $f_{\hat{Z}_{j1}|\lambda_{j1}}$ et la densité de λ_{j1} par $f_{\lambda_{j1}}$. Donc

$$A_{1j1} = \int_{-\infty}^{+\infty} f_{\hat{Z}_{j1}|\lambda_{j1}}\left(\hat{Z}_{j1}\right) f_{\lambda_{j1}}\left(\lambda_{j1}\right) d\lambda_{j1}.$$

$f_{\hat{Z}_{j1}|\lambda_{j1}}\left(\hat{Z}_{j1}\right)$ est donnée par 2.7 et $f_{\lambda_{j1}}\left(\lambda_{j1}\right)$ par 2.4.
Alors

$$\begin{aligned} A_{1j1} &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{\left(1 + \frac{1}{2}\lambda^2_{j1}\right)\sigma^2_{\lambda_{j1}}}} e^{-\frac{1}{2}\left[\frac{\left(\hat{Z}^2_{j1} - \sqrt{N}\lambda_{j1}\right)^2}{1 + \frac{1}{2}\lambda^2_{j1}} + \frac{\lambda^2_{j1}}{\sigma^2_{\lambda_{j1}}}\right]} d\lambda_{j1} \qquad \text{et} \\ A_{2j1} &= f_{\hat{Z}_{j1}|\lambda_{j1}}\left(\hat{Z}_{j1}\right) \quad \text{tel que} \quad \lambda_{j1} = 0 \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{Z}^2_{j1}}{2}}. \end{aligned}$$

À la fin de l'étape 1, nous avons les m probabilités *a posteriori* $q_1 = \{q_{11}, \cdots, q_{m1}\}$. Nous sélectionnons alors le SNP avec la plus grande probabilité *a posteriori*, $q^*_1 = \max(q_1)$. Ensuite

nous changeons les indices de tel sorte que le SNP sélectionné à l'étape 1 soit le SNP 1 codé par X_1 avec sa probabilité *a posteriori* donnée par q_1^*. Après la sélection du premier SNP, nous passons à la sélection du deuxième SNP j pour $j = 2, \cdots, m$.

2.2.2 Étape 2

Nous réindexant les SNPs et leur code de sorte que le SNP 1 est le SNP sélectionné à l'étape 1 et son codage à l'étape 2 est noté par X_{i12} (code du SNP sélectionné à l'étape 1 parmi les m autres codes) pour l'individu i. Considérons les modèles d'hypothèses ci-dessous dont le trait est indépendant du SNP,

$$Y_i = \beta_{0j2} + \beta_{1j2} X_{i12} + \beta_{2j2} X_{ij2} + \varepsilon_{ij2}, \qquad j = 2, \cdots, m$$

où $\varepsilon_{ij2} \sim \mathcal{N}\left(0, \sigma_{j2}^2\right)$, $\sigma_{j2} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape 2 et $\forall i = 1, \cdots, N$.
Posons

$$\tilde{X}_{j2} = \begin{pmatrix} 1 & X_{112} & X_{1j2} \\ \vdots & \vdots & \\ 1 & X_{N12} & X_{Nj2} \end{pmatrix}$$

Nous avons donc

$$\tilde{X}_{j2}^T \tilde{X}_{j2} = N \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V_{j2} \end{pmatrix},$$

avec

$$V_{j2} = \begin{pmatrix} 1 & r_{1j2} \\ r_{1j2} & 1 \end{pmatrix}$$

où $r_{1j2} = N^{-1} \sum_{i=1}^{N} X_{i12} X_{ij2}$ est la corrélation empirique entre les SNPs 1 et j.
Nous avons donc $\left(\tilde{X}_{j2}^T \tilde{X}_{j2}\right)^{-1} = N^{-1} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & W_{j2} \end{pmatrix}$ avec

$$W_{j2} = V_{j2}^{-1} = \frac{1}{1 - r_{1j2}^2} \begin{pmatrix} 1 & -r_{1j2} \\ -r_{1j2} & 1 \end{pmatrix}.$$

Les statistiques de test du type Wald sont donc

$$\hat{Z}_{lj2} = \sqrt{N(1 - r_{1j2}^2)} \frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}_{j2}^2}} \qquad l = 1,2. \tag{2.8}$$

Nous montrons dans l'annexe A.2 que

$$\hat{Z}_{lj2} \sim \mathcal{N}\left(\sqrt{N}\sqrt{1 - r_{1j2}^2}\lambda_{lj2}, 1 + \frac{1}{2}\left(1 - r_{1j2}^2\right)\lambda_{lj2}^2\right), \qquad l = 1,2, \qquad j = 2, \cdots, m. \tag{2.9}$$

Posons $\hat{Z}_{j2} = \left(\hat{Z}_{1j2}, \hat{Z}_{2j2}\right)^T$, nous avons alors

$$\hat{Z}_{j2} \sim \mathcal{N}_2\left(M_{j2}, \Sigma_{j2}\right), \qquad j = 2, \cdots, m \tag{2.10}$$

avec

$$M_{j2} = \sqrt{N}\sqrt{1 - r_{1j2}^2}\lambda_{j2}, \tag{2.11}$$

$$\Sigma_{j2} = \begin{pmatrix} 1 & -r_{1j2} \\ -r_{1j2} & 1 \end{pmatrix} + \frac{1}{2}\left(1 - r_{1j2}^2\right)\lambda_{j2}\lambda_{j2}^T, \tag{2.12}$$

$$\lambda_{j2} = \left(\lambda_{1j2}, \lambda_{2j2}\right)^T. \tag{2.13}$$

Notez que dans les études précédentes qui ignoraient la variabilité générée par l'estimation des termes d'erreurs ont une matrice de variance-covariance de $\hat{Z}_{j2}$ donnée par $\begin{pmatrix} 1 & r_{1j2} \\ r_{1j2} & 1 \end{pmatrix}$.
Nous supposons que les probabilités *a priori* ne dépendent pas des étapes et sont deux à deux indépendantes. Alors $p_1 = \mathbb{P}\left(\lambda_{1j2} = 0\right)$, $p_j = \mathbb{P}\left(\lambda_{2j2} = 0\right)$ et $\mathbb{P}\left(\lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right) = \left(1 - p_1\right)\left(1 - p_j\right)$, pour $j = 2, \cdots, m$.
La probabilité *a posteriori* pour mesurer l'implication du SNP j est donnée par

$$\begin{aligned} q_{j2} &= \mathbb{P}\left(\lambda_{2j2} \neq 0 \mid \hat{Z}_{j2}, \lambda_{1j2} \neq 0\right) \\ &= \frac{\mathbb{P}\left(\lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right)\mathbb{P}\left(\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right)}{\mathbb{P}\left(\lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right)\mathbb{P}\left(\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right) + \mathbb{P}\left(\lambda_{1j2} \neq 0, \lambda_{2j2} = 0\right)\mathbb{P}\left(\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} = 0\right)} \\ &= \frac{(1 - p_1)(1 - p_j)A_{1j2}}{(1 - p_1)(1 - p_j)A_{1j2} + (1 - p_1)p_j A_{2j2}} \end{aligned}$$

où $A_{1j2} = \mathbb{P}\left(\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0\right)$ et $A_{2j2} = \mathbb{P}\left(\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} = 0\right)$.
Pour ne pas alourdir nos calculs, nous considérons que les réalisations des vecteurs de variables aléatoires $\hat{Z}_{j2}$ et λ_{j2} sont aussi notées respectivement par $\hat{Z}_{j2}$ et λ_{j2}. Les réalisations des variables aléatoires λ_{1j2} et λ_{2j2} sont aussi notées respectivement par λ_{1j2} et λ_{2j2}. Nous notons la densité de $\hat{Z}_{j2} \mid \lambda_{1j2} \neq 0, \lambda_{2j2} \neq 0$ par $f_{\hat{Z}_{j1}|\lambda_{j2}}$, la densité de λ_{j2} par $f_{\lambda_{j2}}$, la densité de λ_{1j2} par $f_{\lambda_{1j2}}$ et la densité de λ_{2j2} par $f_{\lambda_{2j2}}$. Donc

$$A_{1j2} = \int_{\mathbb{R}^2} f_{\hat{Z}_{j1}|\lambda_{j2}}\left(\hat{Z}_{j2}\right) f_{\lambda_{j2}}\left(\lambda_{j2}\right) d\lambda_{j2}.$$

Puisque λ_{1j2} et λ_{2j2} sont indépendantes,

$$A_{1j2} = \int_{\mathbb{R}^2} f_{\hat{Z}_{j1}|\lambda_{j2}}\left(\hat{Z}_{j2}\right) f_{\lambda_{1j2}}\left(\lambda_{1j2}\right) f_{\lambda_{2j2}}\left(\lambda_{2j2}\right) d\lambda_{j2}.$$

Comme $f_{\hat{Z}_{j1}|\lambda_{j2}}\left(\hat{Z}_{j2}\right)$ est donnée par 2.10, $f_{\lambda_{1j2}}\left(\lambda_{1j2}\right)$ et $f_{\lambda_{2j2}}\left(\lambda_{2j2}\right)$ par 2.4, alors

$$A_{1j2} = \int_{\mathbb{R}^2} \frac{\sqrt{det\left(\Sigma_{j2}^{-1}\right)}}{(2\pi)^2\sigma_{\lambda_{j2}}^2} \exp\left\{-\frac{1}{2}\left[\left(\hat{Z}_{j2} - M_{j2}\right)^T \Sigma_{j2}^{-1}\left(\hat{Z}_{j2} - M_{j2}\right) + \frac{\lambda_{1j2}^2 + \lambda_{2j2}^2}{\sigma_{\lambda_{j2}}^2}\right]\right\} d\lambda_{j2}.$$

Nous avons aussi

$$A_{2j2} = \int_{-\infty}^{+\infty} f_{\hat{Z}_{j1}|\lambda_{j2}}\left(\hat{Z}_{j2}\right) f_{\lambda_{1j2}}\left(\lambda_{1j2}\right) d\lambda_{1j2} \quad \text{tel que} \quad \lambda_{2j2} = 0.$$

Donc

$$A_{2j2} = \int_{-\infty}^{+\infty} \frac{\sqrt{det\left(\left(\Sigma_{j2}^{0}\right)^{-1}\right)}}{(2\pi)^{3/2}\sigma_{\lambda_{j2}}} \exp\left\{-\frac{1}{2}\left[\left(\hat{Z}_{j2} - M_{j2}^{0}\right)^{T}\left(\Sigma_{j2}^{0}\right)^{-1}\left(\hat{Z}_{j2} - M_{j2}^{0}\right) + \frac{\lambda_{1j2}^{2}}{\sigma_{\lambda_{j2}}^{2}}\right]\right\} d\lambda_{1j2}.$$

Avec M_{j2}^{0} et Σ_{j2}^{0} sont égales à M_{j2} et Σ_{j2} si $\lambda_{2j2} = 0$.
À la fin de l'étape 2, nous avons les $m-1$ probabilités *a posteriori* $q_2 = \{q_{22}, q_{32}, \cdots, q_{m2}\}$. Nous sélectionnons le deuxième SNP avec la plus grande probabilité *a posteriori*, q_2^* est le maximum du vecteur q_2 des probabilités *a posteriori* calculés à l'étape 2. Ensuite nous changeons les indices de telle sorte que le SNP sélectionné à l'étape 2 soit le SNP 2 codé aussi par X_2 et de probabilité *a posteriori* q_2^*. Donc la procédure sélectionne les SNPs 1 et 2 codés par X_1 et X_2 respectivement dans l'ensemble cible Ψ. Après la deuxième étape, nous pouvons passer avec le même principe à la sélection du troisième SNP j pour $j = 3, \cdots, m$. Par la suite, nous supposons que les SNPs $1,2,3, \cdots, k-1$ sont sélectionnés dans l'ensemble cible Ψ par itération avec la méthode décrite ci-dessus. Nous décrivons dans la section suivante comment nous pouvons sélectionner les SNPs k, $k > 2$.

2.2.3 Étape k

À la fin de l'étape $k-1$, l'ensemble cible Ψ contient $k-1$ SNPs indexés 1,...,$k-1$. À l'étape k, on calcule les probabilités *a posteriori* de chacun des $m-k+1$ SNPs restants (indexés k,...,m). Donc les SNPs $1,2, \cdots, k-1$ ont été choisi au $k-1$ premières étapes et leurs codes sont notés par $X_{i1k}, X_{i2k}, \cdots, X_{i,k-1,k}$. Considérons les modèles d'hypothèses ci-dessous dont le trait est indépendant du SNP,

$$Y_i = \beta_{0jk} + \beta_{1jk}X_{i1k} + \cdots + \beta_{k-1,j,k}X_{i,k-1,k} + \beta_{kjk}X_{ijk} + \varepsilon_{ijk}, \qquad j = k, \cdots, m$$

où $\varepsilon_{ijk} \sim \mathcal{N}\left(0, \sigma_{jk}^{2}\right)$, $\sigma_{jk} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape k et $\forall i = 1, \cdots, N$.

Soit $\tilde{X}_{jk} = \begin{pmatrix} 1 & X_{11k} & \cdots & X_{1,k-1,k} & X_{1jk} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & X_{N1k} & \cdots & X_{N,k-1,k} & X_{Njk} \end{pmatrix}$ donc $\tilde{X}_{jk}^{T}\tilde{X}_{jk} = N \begin{pmatrix} 1 & \mathbf{0}^{T} \\ \mathbf{0} & V_{jk} \end{pmatrix}$

et $\left(\tilde{X}_{jk}^{T}\tilde{X}_{jk}\right)^{-1} = N^{-1} \begin{pmatrix} 1 & \mathbf{0}^{T} \\ \mathbf{0} & W_{jk} \end{pmatrix}$ et $W_{jk} = V_{jk}^{-1}$.

Soit W_{jklq} l'élément à la position (l,q) de la matrice W_{jk}.
La statistique d'association peut être écrite comme suit

$$\hat{Z}_{ljk} = \frac{\sqrt{N}}{\sqrt{W_{jkll}}} \frac{\hat{\beta}_{ljk}}{\sqrt{\hat{\sigma}_{jk}^{2}}} \qquad l = 1,2, \cdots, k.$$

Nous montrons dans l'annexe A.3 que

$$\hat{Z}_{ljk} \sim \mathcal{N}\left(\frac{\sqrt{N}}{\sqrt{W_{jkll}}}\lambda_{ljk}, 1 + \frac{1}{2}\frac{1}{W_{jkll}}\lambda_{ljk}^2\right). \tag{2.14}$$

Considérons le vecteur des statistiques d'association noté par $\hat{Z}_{jk} = \left(\hat{Z}_{1jk}, \hat{Z}_{2jk}, \cdots, \hat{Z}_{kjk}\right)^T$.
Nous montrons que

$$\hat{Z}_{jk} \sim \mathcal{N}_k\left(M_{jk}, \Sigma_{jk}\right) \tag{2.15}$$

avec $M_{jk} = D_{jk}\lambda_{jk}$ où D_{jk} est une matrice diagonale de dimensions $k \times k$ et $\lambda_{jk} = (\lambda_{1jk}, \lambda_{2jk}, \cdots, \lambda_{k-1,j,k}, \lambda_{kjk})^T$,

$$D_{jk} = \begin{pmatrix} \frac{\sqrt{N}}{\sqrt{W_{jk11}}} & 0 & \cdots & 0 \\ 0 & \frac{\sqrt{N}}{\sqrt{W_{jk22}}} & & \vdots \\ \vdots & & \ddots & \\ 0 & & & \frac{\sqrt{N}}{\sqrt{W_{jkkk}}} \end{pmatrix}$$

et

$$\Sigma_{jk} = R_{jk}^1 + \frac{1}{2}R_{jk}^2\lambda_{jk}^T\lambda_{jk}$$

avec R_{jk}^1 et R_{jk}^2 sont des matrices de dimensions $k \times k$ telles que

$$R_{jk}^1 = \begin{pmatrix} 1 & \frac{W_{jk12}}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \cdots & \frac{W_{jk1k}}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} \\ \frac{W_{jk12}}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & 1 & & \\ \vdots & & \ddots & \\ \frac{W_{jk1k}}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} & & & 1 \end{pmatrix},$$

$$R_{jk}^2 = \begin{pmatrix} \frac{1}{W_{jk11}} & \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \cdots & \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} \\ \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \frac{1}{W_{jk11}} & & \\ \vdots & & \ddots & \\ \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} & & & \frac{1}{W_{jkkk}} \end{pmatrix}.$$

Notez que $R_{jk}^2\lambda_{jk}^T\lambda_{jk}/2$ est un terme de correction de la matrice de variance-covariance de Σ_{jk}.
La probabilité *a posteriori* pour mesurer l'implication du SNP j est donnée par

$$\begin{aligned} q_{jk} &= \mathbb{P}\left(\lambda_{kjk} \neq 0 \mid \hat{Z}_{jk}, \lambda_{1jk} \neq 0, \lambda_{2jk} \neq 0, \cdots, \lambda_{k-1,j,k} \neq 0\right) \\ &= \frac{(1-p_1)(1-p_2)\cdots(1-p_j)A_{1jk}}{(1-p_1)(1-p_2)\cdots(1-p_j)A_{1jk} + (1-p_1)(1-p_2)\cdots(1-p_{k-1})p_jA_{2jk}} \end{aligned}$$

où $A_{1jk} = \mathbb{P}\left(\hat{Z}_{jk} \mid \lambda_{1jk} \neq 0, \lambda_{2jk} \neq 0, \cdots, \lambda_{k-1,j,k} \neq 0, \lambda_{kjk} \neq 0\right)$ et $A_{2jk} = \mathbb{P}\left(\hat{Z}_{jk} \mid \lambda_{1jk} \neq 0, \lambda_{2jk} \neq 0, \cdots, \lambda_{k-1,j}\right.$
Pour ne pas alourdir nos calculs, nous considérons que les réalisations des vecteurs de variables aléatoires $\hat{Z}_{jk}$ et λ_{jk} sont aussi notées respectivement par $\hat{Z}_{jk}$ et λ_{jk}. Les réalisations

des variables aléatoires λ_{1jk},λ_{2jk},$\cdots$,$\lambda_{k-1,j,k}$ et λ_{kjk} sont aussi notées respectivement par λ_{1jk},λ_{2jk},$\cdots$,$\lambda_{k-1,j,k}$ et λ_{kjk}. Nous notons la densité de $\hat{Z}_{jk} \mid \lambda_{1jk} \neq 0, \lambda_{2jk} \neq 0, \cdots, \lambda_{k-1,j,k} \neq 0, \lambda_{jjk} \neq 0$ par $f_{\hat{Z}_{jk}|\lambda_{jk}}$, la densité de λ_{jk} par $f_{\lambda_{jk}}$ et les densités de λ_{1jk},λ_{2jk},$\cdots$,$\lambda_{k-1,j,k}$, λ_{kjk} par $f_{\lambda_{1jk}}$,$f_{\lambda_{2jk}}$,$\cdots$,$f_{\lambda_{k-1,j,k}}$, $f_{\lambda_{kjk}}$ respectivement. Donc

$$A_{1jk} = \int_{\mathbb{R}^k} f_{\hat{Z}_{jk}|\lambda_{jk}}\left(\hat{Z}_{jk}\right) f_{\lambda_{jk}}\left(\lambda_{jk}\right) d\lambda_{jk}.$$

Puisque λ_{1jk},λ_{2jk},$\cdots$,$\lambda_{k-1,j,k}$ et λ_{kjk} sont deux à deux indépendantes, alors

$$A_{1jk} = \int_{\mathbb{R}^k} f_{\hat{Z}_{jk}|\lambda_{jk}}\left(\hat{Z}_{jk}\right) f_{\lambda_{1jk}}\left(\lambda_{1jk}\right) \cdots f_{\lambda_{kjk}}\left(\lambda_{kjk}\right) d\lambda_{jk}.$$

Nous obtenons $f_{\hat{Z}_{jk}|\lambda_{jk}}\left(\hat{Z}_{jk}\right)$ par 2.15 et $f_{\lambda_{1jk}}\left(\lambda_{1jk}\right) \cdots f_{\lambda_{kjk}}\left(\lambda_{kjk}\right)$ par 2.4. D'où

$$A_{1jk} = \int_{\mathbb{R}^k} \frac{\sqrt{det\left(\Sigma_{jk}^{-1}\right)}}{(2\pi)^k \sigma_{\lambda_{jk}}^k} \exp\left\{-\frac{1}{2}\left[\left(\hat{Z}_{jk} - M_{jk}\right)^T \Sigma_{jk}^{-1}\left(\hat{Z}_{jk} - M_{jk}\right) + \frac{\lambda_{1j2}^2 + \cdots + \lambda_{kjk}^2}{\sigma_{\lambda_{jk}}^2}\right]\right\} d\lambda_{jk}.$$

De même, nous avons

$$A_{2jk} = \int_{\mathbb{R}^{k-1}} f_{\hat{Z}_{jk}|\lambda_{jk}}\left(\hat{Z}_{jk}\right) f_{\lambda_{1jk}}\left(\lambda_{1jk}\right) \cdots f_{\lambda_{k-1,j,k}}\left(\lambda_{k-1,j,k}\right) d\lambda_{jk} \quad \text{tel que} \quad \lambda_{kjk} = 0.$$

Par conséquent

$$\begin{aligned} A_{2jk} &= \int_{\mathbb{R}^{k-1}} \frac{\sqrt{det\left(\left(\Sigma_{jk}^0\right)^{-1}\right)}}{(2\pi)^{(2k-1)/2} \sigma_{\lambda_{jk}}^{k-1}} \exp\left\{-\frac{1}{2}\left[\left(\hat{Z}_{jk} - M_{jk}^0\right)^T \left(\Sigma_{jk}^0\right)^{-1}\left(\hat{Z}_{jk} - M_{jk}^0\right) + \frac{\lambda_{1j2}^2 + \cdots + \lambda_{k-1,j,k}^2}{\sigma_{\lambda_{jk}}^2}\right]\right\} \\ &\times \quad d\lambda_{1j2} \cdots d\lambda_{k-1,j,k} \end{aligned} \tag{2.17}$$

Avec M_{jk}^0 et Σ_{jk}^0 sont égales à M_{jk} et Σ_{jk} si $\lambda_{kjk} = 0$.
À la fin de l'étape k, nous avons les $m-k+1$ probabilités *a posteriori* $q_k = \{q_{kk}, q_{k+1,k}, \cdots, q_{mk}\}$. Nous sélectionnons le k-ème SNP avec la plus grande probabilité *a posteriori*, $q_k^* = \max(q_k)$. Ensuite nous changeons les indices de telle sorte que le SNP sélectionné à l'étape k soit le SNP k codé aussi par X_k dont la probabilité *a posteriori* d'être significative est q_k^*. Donc la procédure sélectionne les SNPs 1,2,$\cdots$,k codés par $X_1, X_2, \cdots, X_k$ respectivement dans l'ensemble cible Ψ. Les intégrales A_{1jk} et A_{2jk} seront calculées par approximation de Monte Carlo (voir la section 3.2) dans nos simulations.

2.2.4 Résumé de la méthode de sélection de SNP

La procédure de sélection de SNPs présentée dans ce mémoire est une méthode conditionnelle. Nous avons m SNPs pour N individus dont le trait d'intérêt Y est mesuré et les probabilités *a priori* 2.3 sur l'état causal de chaque SNP. Notre objectif est de sélectionner K SNPs les plus significativement associés à Y parmi les m SNPs de la base de données. Soit le code des m SNPs noté par $(G_1, \cdots, G_m)$, nous normalisons ces valeurs par $(X_1, \cdots, X_m)$. Nous avons donc la base de données $\{Y; \quad X_1, \cdots, X_m; \quad p_1, \cdots, p_m\}$. Notre méthode est itérative. Elle calcule à chaque étape la statistique d'association des SNPs et leurs probabilités *a posteriori* 2.5 en utilisant la loi *a priori* définie en 2.4. À chaque étape, le SNP avec la plus forte probabilité *a posteriori* est choisi.

Ainsi à l'étape 1, nous avons la base notée par $\{Y; \quad X_1, \cdots, X_m; \quad p_1, \cdots, p_m\}$. Nous calculons d'abord les statistiques d'association 2.6 et leur loi est donnée par 2.7. Ensuite les probabilités *a posteriori* $q_{11}, \cdots, q_{m1}$ des SNPs sont calculées. Nous sélectionnons le premier SNP tel que sa probabilité *a posteriori* est $q_1^* = \max_{1 \leq j \leq m} (q_{j1})$. Avant de passer à l'étape 2, nous réindexons les codes des SNPs tel que le SNP choisi est le SNP 1 et est noté par $X_1 \equiv X_1^c$.

À l'étape 2, nous intégrons dans le modèle le SNP 1 sélectionné à l'étape 1 et nous considérons la base de données restreinte $\{Y; \quad X_2, \cdots, X_m; \quad p_1, \cdots, p_m\}$. Ensuite nous calculons les statistiques d'association 2.8 des SNPs de la base de données restreinte. Enfin, nous calculons les probabilités *a posteriori* $q_{22}, \cdots, q_{m2}$ et le deuxième SNP est sélectionné tel que sa probabilité *a posteriori* vérifie $q_2^* = \max_{2 \leq j \leq m} (q_{j2})$. Nous réindexons les codes des SNPs tel que le deuxième SNP choisi est le SNP 2 et est noté par $X_2 \equiv X_2^c$.

Notons que grâce à la loi *a priori* 2.4, les probabilités *a posteriori* qui nous permettent de sélectionner les SNPs dépendent d'intégrales qui peuvent être facilement calculées en utilisant des approximations de Laplace. Ceci nous permet d'optimiser le temps de calcul des probabilités *a posteriori* qui est souvent fastidieux [27].

Par itération, nous sélectionnons les K SNPs de l'ensemble cible Ψ cherché, $\Psi = (X_1^c, \cdots, X_K^c)$ qui ont une association significative avec le trait d'intérêt Y. À la fin, notre procédure sélectionne les SNPs $\{X_1^c, \cdots, X_K^c\}$. Par la suite des techniques génétiques permettent de déterminer les gènes associés aux SNPs sélectionnés et le biologiste peut valider l'association fonctionnelle des K SNPs au trait d'intérêt Y au laboratoire.

Chapitre 3

Simulations

3.1 Paramètres de la distribution de la statistique d'association

3.1.1 Illustration 1

Dans cette section, nous illustrons le fait que la distribution asymptotique trouvée en 2.7 est la bonne distribution pour la statistique d'association définie en 2.6.
Considérons le SNP 1 codé par X_1 et le modèle de régréssion associé à ce SNP,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \qquad i = 1, \cdots, N,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma_1^2)$.
Pour $N = 1000$, $\beta_0 = 1$, $\beta_1 = 2$ et $\sigma_1 = 1$, nous calculons la statistique d'association. Avec 1000 répétitions, nous obtenons le vecteur des statistiques d'association $\hat{Z}_1$ dont l'histogramme est ajusté par la distribution obtenue en 2.7. Nous avons le graphe de l'histogramme de la statistique d'association ci-dessous

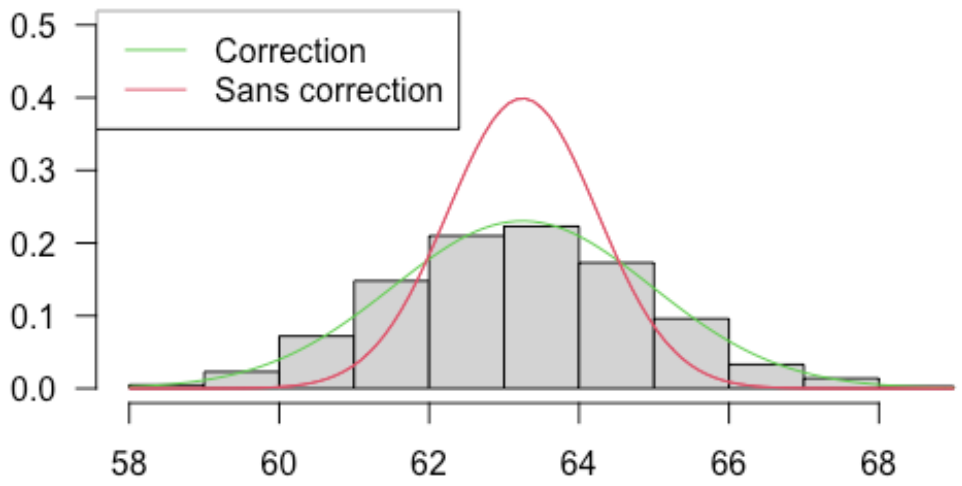

FIGURE 3.1 – Ajustement de la distribution de $\hat{Z}_1$

Dans nos simulations, la variance empirique de $\hat{Z}_1$ est $Var\left(\hat{Z}_1\right) = 2.928394$. Tandis que la variance théorique trouvée dans ce mémoire est

$$1 + \frac{1}{2}\lambda_1^2 = 1 + \frac{1}{2}\left(\frac{\beta_1}{\sigma_1}\right)^2 = 1 + \frac{1}{2}\left(\frac{2}{1}\right)^2 = 3.$$

Alors cet exemple montre bien que la variance de $\hat{Z}_1$ n'est pas égale à 1. Nous avons une bonne approximation de la variance de $\hat{Z}_1$ par la variance théorique trouvée en 2.7. La figure 3.1 montre que la distribution de $\hat{Z}_1$ est bien ajustée par la distribution 2.7 trouvée dans ce mémoire.

3.1.2 Illustration 2

Dans cette illustration, nous voulons vérifier qu'à l'étape 2 la covariance entre deux statistiques d'association est donnée par 2.12 et n'est pas égale à la corrération entre les deux SNPs comme obtenue dans [27], [28].
Grâce à l'algorithme défini en 3.3, nous simulons le code deux SNPs 1 et 2 pour 1000 individus. Ensuite, nous notons les valeurs centrées et réduites de ces deux codes par X_1 et X_2.
Considérons le modèle de régréssion suivant

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad i = 1, \cdots, N,$$

avec $\varepsilon \sim \mathcal{N}(0,\sigma^2)$.
Avec 1000 répétitions, nous calculons les statistiques d'association $\hat{Z}_1$ et $\hat{Z}_2$ de ces deux SNPs en utilisant 2.9. Après nous calculons la covariance entre $\hat{Z}_1$ et $\hat{Z}_2$. Nous répétons ce processus de calcule de la covariance empirique 1000 fois. Donc nous obtenons 1000 covariances entre $\hat{Z}_1$ et $\hat{Z}_2$ dont le diagramme en boîte est donné par la figure 3.2. À la fin, nous ajoutons au diagramme en boîte la ligne horizontale rouge qui représente la covariance théorique entre $\hat{Z}_1$ et $\hat{Z}_2$ trouvée dans ce mémoire :

$$\begin{aligned} Cov\left(\hat{Z}_1, \hat{Z}_2\right) &= -r + \frac{1}{2}\left(1 - r^2\right)\lambda_1\lambda_2 \\ &= 0.1050029 \end{aligned}$$

avec $\lambda_1 = \beta_1/\sigma = 1/1 = 1$, $\lambda_2 = \beta_2/\sigma = 1/1 = 1$ et r la corrélation empirique entre X_1 et X_2, $r = 0.3375755$.

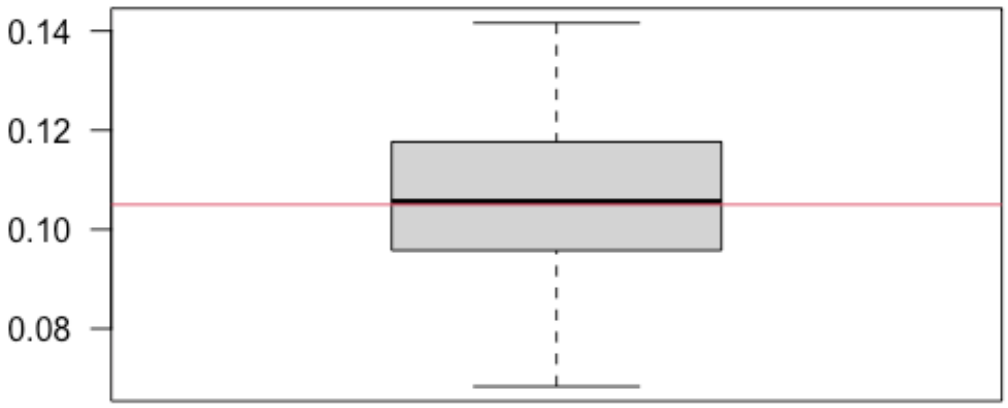

FIGURE 3.2 – Diagramme en boîte de la covariance entre $\hat{Z}_1$ et $\hat{Z}_2$

Donc la figure 3.2 montre que les 1000 covariances entre $\hat{Z}_1$ et $\hat{Z}_2$ sont centrées sur la covariance théorique trouvée dans ce travail. Dans d'autres études [27], [28], la covariance entre $\hat{Z}_1$ et $\hat{Z}_2$ est prise comme étant égale à r. Cependant cette illustration nous montre que la covariance entre $\hat{Z}_1$ et $\hat{Z}_2$ ne peut pas être égale à r.

3.2 Approximation Monte Carlo de A_{1jk} et A_{2jk}

Nous montrons ici qu'une approximation de Monte Carlo est adéquate pour le type de fonction à intégrer dans ce mémoire.

Example 1

Considérons les deux variables aléatoires indépendantes λ_1 et λ_2 de loi $\mathcal{N}(0,\sigma_\lambda)$.
Nous voulons calculer l'intégrale double A_1 définie par

$$A_1 = \int_{\mathbb{R}^2} \lambda_1^2 \lambda_2^2 f(\lambda_1) f(\lambda_2) d\lambda_1 d\lambda_2.$$

Par un calcul direct nous obtenons facilement $A_{1,Exact} = \sigma_\lambda^4$. Numériquement nous pouvons calculer A_1 par approximation de Monte Carlo, $A_{1,MC}$ ou par la fonction *adaptIntegrate()* du package de R nommé *cubature*, $A_{1,CUABAT}$. La fonction *adaptIntegrate()* effectue l'intégration multidimensionnelle adaptative à valeurs vectorielles sur des hypercubes.

Simulation

Posons $\sigma_\lambda = 2$ et calculons A_1 par Monte Carlo, $A_{1,MC}$ et par la fonction *adaptIntegrate()* de *cubature*, $A_{1,CUBAT}$ avec un nombre de répétition $NB = 1000$. Nous avons $A_{1,MC} = 16.01711$, $A_{1,CUBAT} = 16.00002$ et par calcul direct $A_{1,Exact} = 16$.
Donc les deux méthodes d'approximations numériques sont utiles pour l'approximation des types d'intégrales que nous utilisons dans notre mémoire.

Example 2

Considérons le couple de variables aléatoires $Z = (Z_1,Z_2)$ de loi conditionnelle

$$Z \mid \lambda \sim \mathcal{N}\left(\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}, \begin{pmatrix}1 & r_{12}\\ r_{12} & 1\end{pmatrix}\right)$$

tel que $\lambda = (\lambda_1,\lambda_2)$

$$\lambda \sim \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_{\lambda_1} & 0\\ 0 & \sigma_{\lambda_2}\end{pmatrix}\right).$$

Nous voulons maintenant calculer l'intégrale

$$A_2 = \int_{\mathbb{R}^2} f_Z(x_1,x_2) f_{\lambda_1}(y_1) f_{\lambda_2}(y_2) dy_1 dy_2$$

avec x_1, x_2, y_2 et y_2 les réalisations des variables aléatoires Z_1, Z_2, λ_1 et λ_2 respectivement.
Trouvons d'abord la loi exacte de Z.
Proposition : *Conditionnellement à la variable aléatoire m, soit X une variable aléatoire de*

loi normale $\mathcal{N}(m,\sigma^2)$*, avec* $\sigma > 0$*. Nous supposons que* m *suit une loi normale* $\mathcal{N}(\theta,s^2)$*, avec* $\theta \in \mathbb{R}$ *et* $s > 0$*. Alors la loi exacte de* X *est donnée par*

$$X \sim \mathcal{N}\left(\theta,s^2+\sigma^2\right).$$

Nous pouvons donc déterminer la loi exacte du couple par

$$Z \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1+\sigma_{\lambda_1^2} & r_{12}\\ r_{12} & 1+\sigma_{\lambda_2^2}\end{pmatrix}\right)$$

Par un calcul direct, nous avons $A_{2,Exact} = f_Z(x_1,x_2)$. Numériquement nous pouvons aussi calculer A_2 par approximation de Monte Carlo, $A_{2,MC}$ ou par la fonction *adaptIntegrate()* du package de R nommé *cubature*, $A_{2,CUBAT}$.

Simulation

Pour différentes valeurs des paramètres r_{12}, λ et σ_λ, nous simulons 1000 valeurs du vecteur Z dont la loi exacte est obtenue ci-dessus. Nous voulons calculer

$$A_2 = \int_{\mathbb{R}^2} f_Z(x_1,x_2)\, f_{\lambda_1}(y_1)\, f_{\lambda_2}(y_2)\, dy_1 dy_2$$

Par calcul direct avec chaque réalisatin de Z, nous calculons $f_Z(x_1,x_2)$ et $A_{2,Exact}$ est la moyenne du résultat trouvée.
Par la méthode de Monte Carlo, nous calculons A_2 en utilisant la loi de $Z \mid \lambda$. Pour cela, nous simulons 1000 variables alátoires $\lambda \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{\lambda_1} & 0\\ 0 & \sigma_{\lambda_2}\end{pmatrix}\right)$. Ensuite, nous calculons la densité du vecteur Z pour réalisation de la variable aléatoire λ. D'après Monte Carlo, $A_{2,Mc}$ est la moyenne des densités de Z.
Par la méthode des curbature, nous calculons la fonction de densité de Z en fonction de λ. $A_{2,CUBAT}$ est la valeur de la fonction *adaptIntegrate()* dont la fonction à intégrer est fonction de densité de Z en fonction de λ avec comme borne inférieur, $\lambda = -\infty$ et comme borne supérieure, $\lambda = +\infty$. De plus nous laissons autres paramètres de *adaptIntegrate()* par défaut. Nous avons le tableau ci-dessous des approximations de A_2.

r_{12}	λ	σ_λ	$A_{2,Exact}$	$A_{2,CUBAT}$	$A_{2,MC}$
0.5	(1, 1)	(1, 1)	0.03142625	0.03142626	0.03232844
0.5	(3, 2)	(0.1, 0.4)	0.002141902	0.002241659	0.002139709
0.8	(0.1, 0.4)	(0.5, 0.1)	0.1430748	0.1307301	0.1432787

Dans cette section, nous nous limitons aux calculs d'intégrales simple et double pour choisir une méthode d'approximation des nos intégrales obtenues dans ce travail. En dimension deux, la méthode de Monte Carlo est adaptée pour la forme de fonction que nous intégrons dans ce

mémoire. Nous avons un gain de temps en utilisant une approximation de Monte Carlo. Avec 1000 répétions, nous calculons le temps de calcul de A_2 (avec la fonction *system.time()*) pour la méthode de Monte Carlo et la méthode de cubature. Nous avons

Temps	utilisateur	système	écoulé
Monte Carlo	0.000	0.000	0.001
Cubature	0.228	0.016	0.248

Pour optimiser notre procédure de sélection de SNPs, nous choisissons d'utiliser la méthode de Monte Carlo dans ce mémoire. Nous voulons maintenant déterminer le nombre de répétion B nécessaire à la méthode de Monte Carlo pour avoir une bonne approximation de l'intégrale.

Choix de B dans l'approximation de Monte Carlo

Dans les mêmes conditions que l'exemple 2 (avec toujours 1000 répétitions), nous calculons les intgrales pour des différentes valeurs de B. Nous avons

B	10	100	500	1000	1500	10000
$A_{2,Exact}$	0.1765994	0.1765994	0.1765994	0.1765994	0.1765994	0.1765994
$A_{2,MC}$	0.1845388	0.1872712	0.1815981	0.1795753	0.1737860	0.1758343

Dans les simulations de ce mémoire nous utilisons par défaut B=1000 pour calculer les intégrales A_{1jk} et A_{2jk} dans notre procédure de sélection de SNPs.

3.3 Méthode de génération de données de type SNP

Dans ce mémoire, nous définissons le code d'un SNP comme étant le nombre d'allèles mineurs présents sur la paire d'allèles représentant ce SNP. Donc le code d'un SNP prend ici les valeurs 0 si le SNP n'a pas d'allèles mineures, 1 s'il a une seule allèle mineure et 2 s'il en a deux. Nous générons des données de type SNP pour étudier notre procédure de sélection de SNPs. La fonction $SNP012()$ (voir annexe B.1) dont l'algorithme est défini ci-dessous, nous permet de générer des données de type SNP. Cette fonction retourne un tableau de données dont les éléments sont 0 ou 1 ou 2 qui représentent le nombre d'allèles mineures d'un SNP. Le SNP j sera codé par G_j. Soit le SNP j qui a deux locus bi-alléliques notés L_{j1} et L_{j2}.
Considérons les deux SNPs 1 et 2 formés par les deux paires d'allèles a/A et b/B. Les allèles mineures sont notées par a et b de fréquence allélique f_1 et f_2 respectivement, alors

$$\mathbb{P}(L_{11} = a) = f_1, \qquad \mathbb{P}(L_{12} = A) = 1 - f_1,$$
$$\mathbb{P}(L_{21} = b) = f_2, \quad \text{et} \quad \mathbb{P}(L_{22} = B) = 1 - f_2.$$

Nous pouvons coder les deux SNPs 1 et 2 additivement comme 0, 1 ou 2 pour le nombre d'allèles mineures,

$$G_1 = \mathbb{1}_{\{L_{11}=a\}} + \mathbb{1}_{\{L_{12}=a\}} \qquad \text{et} \qquad G_2 = \mathbb{1}_{\{L_{21}=b\}} + \mathbb{1}_{\{L_{22}=b\}}.$$

Donc nous avons $G_j \in \{0,1,2\}$, $j = 1,2$.
La fonction de génération de SNPs prend comme paramètres le nombre d'individus N, la fréquence de l'allèle mineur de chaque SNP $0 \leq f \leq 0.5$ et la matrice de corrélation V entre les SNPs. Elle génère des données de type SNP.
Nous générons d'abord des variables aléatoires Z à partir d'une distribution normale multivariée de moyennes nulles et de variance covariance V. Par discrétisation avec les quantiles de la distribution normale, nous avons la variable aléatoire $U = \Phi(Z)$.
Soient a une allèle mineur de fréquence allélique f et N_a le nombre d'allèles mineurs d'un SNP codé par G. Si l'équilibre de Hardy-Weinberg 1.1.4 est vérifié alors $\mathbb{P}(N_a = 0) = (1-f)^2$, $\mathbb{P}(N_a = 1) = 2f(1-f)$ et $\mathbb{P}(N_a = 2) = f^2$. Ces probabilités nous aident à établir l'algorithme qui nous permet de générer des données de type SNP.

Algorithme

- si $U \leq (1-f)^2$ alors $G = 0$,
- si $(1-f)^2 < U < (1-f)^2 + 2f(1-f)$ alors $G = 1$,
- si $U \geq (1-f)^2 + 2f(1-f)$ alors $G = 2$.

Illustration

Avec la fonction décrite ci-dessous, nous générons des codes pour six SNPs.
Paramètres : $N = 1500$, $f = (0.01, 0.1, 0.2, 0.3, 0.4, 0.5)$ et

$$V_G = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

Donc nous avons le code des six SNPs pour les six premiers individus présenté dans le tableau ci-dessous.

Individu	G_1	G_2	G_3	G_4	G_5	G_6
1	0	0	0	1	1	0
2	0	1	1	0	2	2
3	0	0	0	1	0	0
4	0	0	0	0	0	0
5	0	0	0	1	1	1
6	0	0	1	0	2	1

Nous pouvons centrer et réduire les SNPs G_j par $X_j = \left(G_j - \overline{G}_j\right) / \sqrt{S^2_{G_j}}$ avec $\overline{G}_j = N^{-1} \sum_{k=1}^{N} G_{ik}$ et $S^2_{G_j} = N^{-1} \sum_{i=1}^{N} \left(G_{ji} - \overline{G}_j\right)^2$.

Individu	X_1	X_2	X_3	X_4	X_5	X_6
1	-0.1378733	-0.4591752	-0.6950903	0.6106462	0.3223935	-1.393342187
2	-0.1378733	1.8756138	1.1502821	-0.9418441	1.7789904	1.402662201
3	-0.1378733	-0.4591752	-0.6950903	0.6106462	-1.1342035	-1.393342187
4	-0.1378733	-0.4591752	-0.6950903	-0.9418441	-1.1342035	-1.393342187
5	-0.1378733	-0.4591752	-0.6950903	0.6106462	0.3223935	0.004660007
6	-0.1378733	-0.4591752	1.1502821	-0.9418441	1.7789904	0.004660007

Nous notons qu'avec notre méthode de simulation de SNP, la corrélation V_G entre les SNP fournie à l'entrée de la fonction $SNP012(N,f,V_G)$ est sous estimée par V_X du fait que nous discrétisons par les quantiles de la loi normale. En effet, le code des SNPs est obtenue en utilisant la variable alátoire U. Comme nous discrétisons $Z \sim \mathcal{N}(0,V_G)$ par la variable U, nous perdons des information sur la matrice V_G. Ceci qui peut expliquer une sous estimations de la matrice corrélation des SNPs codés dans ce travail. Nous avons

$$V_X = \begin{pmatrix} 1 & 0.21 & 0.16 & 0.13 & 0.17 & 0.15 \\ 0.21 & 1 & 0.29 & 0.33 & 0.30 & 0.32 \\ 0.16 & 0.29 & 1 & 0.35 & 0.36 & 0.36 \\ 0.13 & 0.33 & 0.35 & 1 & 0.40 & 0.41 \\ 0.17 & 0.30 & 0.36 & 0.40 & 1 & 0.41 \\ 0.15 & 0.32 & 0.36 & 0.41 & 0.41 & 1 \end{pmatrix}$$

3.4 Illustration de la procédure de sélection

Dans cette section nous illustrons le fonctionnement de notre procédure de sélection de SNPs sur des exemples. Nous nous intéressons à la sélection de cinq SNPs, $K = 5$ dans notre ensemble cible Ψ parmi les m SNPs générés par la fonction $SNP012()$. Ensuite nous appliquons la procédure de sélection proposée au chapitre 2.

Dans cette illustration, nous montrons comment notre procédure fonctionne et exposons les résultats qu'elle retourne.

Paramètres :

$N = 1500$, $m = 6$ et la matrice de corrélation V est définie pour une corrélation de 0.5 entre tous les SNPs. La fréquence des allèles mineurs f est donnée $f = (0.14, 0.39, 0.20, 0.44, 0.47, 0.02)$. Le trait d'intérêt Y est supposé associé uniquement aux SNPs $(2, 3, 4, 5, 6)$ tel que l'écart-type de l'erreur aléatoire normale, $\sigma = 1$ et les paramètres sont $\beta_1 = 0$ et $\beta_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0.01$.

Les information *a priori* :

Les probabilités *a priori* (les probabilités qu'un SNP soit causal), $1 - p_2 = 1 - p_4 = 0.6$, $1 - p_3 = 1 - p_6 = 0.7$, $1 - p_5 = 0.9$ et $1 - p_1 = 0.2$.

Les écart-types sur la loi *a priori* de λ définie en 2.4 est $\sigma_{\lambda_j} = 0.1$, $j = 1, \cdots, m$.

* **Étape 1** : La procédure sélectionne parmi les SNPs associés au trait Y, le SNP 5 dont la probabilité *a posteriori* est plus forte, $q_{51} = 0.9522815$.

SNP j	1	2	3	4	5	6
$1 - p_j$	0.2	0.6	0.7	0.6	0.9	0.7
$\hat{Z}_{j1}$	2.17	1.85	1.02	1.14	2.15	1.09
$\hat{\lambda}_{j1}$	0.06	0.05	0.03	0.03	0.05	0.03
q_{j1}	0.3738744	0.6498613	0.4857998	0.4091988	0.9522815	0.4951067

* **Étape 2** : Le vecteur des statistiques d'associations $\hat{Z}_{j2}$ est constitué en première position de la statistique d'association correspondant au paramètre du SNP sélectionné à l'étape 1 et en seconde position de la statistique d'association du SNP j. De même, le paramètre est un vecteur de dimension 2. Le premier élément du vecteur $\hat{\lambda}_{j2}$ correspond au paramètre du SNP sélectionné à l'étape 1 et le deuxième élément de ce vecteur est le paramètre du SNP j.

SNP j	1	2	3	4	6
$1 - p_j$	0.2	0.6	0.7	0.6	0.7
$\hat{Z}_{j2}$	(1.43, 1.46)	(1.54, 1.08)	(1.9, 0.18)	(1.86, 0.32)	(1.89, 0.35)
$\hat{\lambda}_{j2}$	(0.04, 0.04)	(0.04, 0.03)	(0.05, 0.01)	(0.05, 0.01)	(0.05, 0.01)
q_{j2}	0.1514521	0.4383936	0.4170020	0.3298301	0.4204794

À l'étape 2, le SNP 2 a la plus grande probabilité *a posteriori* $q_{21} = 0.4383936$. Donc, la procédure s'electionne le SNP 2 comme le plus associé au trait.

* **Étape 3** : Le SNP 3 est sélectionné par la procédure de sélection de SNPs car il a la plus forte probabilité *a posteriori* $q_{33} = 0.4258074$.

SNP j	1	3	4	6
$1 - p_j$	0.2	0.7	0.6	0.7
$\hat{Z}_{j3}$	(1.15, 0.73, 1.21)	(1.52, 1.08, -0.14)	(1.49, 1.04, -0.0009)	(1.45, 1.04, 0.14)
$\hat{\lambda}_{j3}$	(0.03, 0.02, 3.52×10^{-2})	(0.04, 0.03, -4.05×10^{-3})	(0.04, 0.03, -2.64×10^{-5})	(0.04, 0.03, 4.09×10^{-3})
q_{j3}	0.1291313	0.4258074	0.3039819	0.3977920

* **Étape 4** : La procédure sélectionne le SNP 6 avec la plus forte probabilité *a posteriori* $q_{64} = 0.3962411$.

SNP j	1	4	6
$1 - p_j$	0.2	0.6	0.7
$\hat{Z}_{j4}$	(1.20, 0.80, -0.38, 1.27)	(1.48, 1.04, -0.14, 0.03)	(1.45, 1.05, -0.17, 0.18)
$\hat{\lambda}_{j4}$	(0.04, 0.02, -0.01, 0.04)	(0.04, 0.03, -0.004, 0.001)	(0.04, 0.03, -0.01, 0.01)
q_{j4}	0.1241160	0.2727047	0.3962411

* **Étape 5** : D'après le modèle de génération du trait d'intérêt Y, il nous reste le SNP 4 associé à Y et le SNP 1 non associé Y. La procédure sélectionne le SNP 4 car ayant la plus forte probabilité *a posteriori* $q_{45} = 0.26235500$.

SNP j	1	4
$1 - p_j$	0.2	0.6
$\hat{Z}_{j5}$	(1.18, 0.79, -0.38, 0.03, 1.25)	(1.43, 1.02, -0.17, 0.17, 0.01)
$\hat{\lambda}_{j5}$	(0.04, 0.02, -0.01, 0.0009, 0.04)	(0.04, 0.03, -0.01, 0.01, 0.0004)
q_{j5}	0.09992528	0.26235500

À titre d'illustration, nous fixons le nombre de SNPs à sélectionner $K = 5$ avec notre procédure. Le choix de K nécessite une étude dont nous n'aborderons pas dans ce mémoire. Pour plus d'idées dans le choix de K voir [27].
Donc, notre procédure de sélection de SNPs donne l'ensemble cible $\Psi = \{5, 2, 3, 6, 4\}$. Cet ensemble est conforme au modèle de génération du trait Y. Cet ensemble choisi sera utilisé pour les tests au laboratoire pour déterminer leur association fonctionnelle sur le trait Y. Les procédures de sélection de SNPs ont pour objectif de réduire l'ensemble de m SNPs à tester au laboratoire avec l'ensemble cible Ψ.
Après avoir montré les résultats que donne la procédure, nous illustrons ci-dessous la relation qu'il y a entre la probabilité *a priori* donné et le SNP sélectionné. Pour cela, nous considérons toujours les paramètres ci-dessous pour le modèle. Nous regardons le comportement de la procédure de sélection avec des informations *a priori* différentes. Nous obtenons les deux tableaux représentant l'état de sélection d'un SNP. Si la procédure sélectionne un SNP nous notons dans le tableau oui sinon nous notons par non. Les SNPs sont affichés selon l'ordre de sélection.

SNP j	2	5	3	6	4	1
$1 - p_j$	0.5	0.5	0.5	0.5	0.5	0.5
SNP sélectionné (oui/non)	oui	oui	oui	oui	oui	non

Malgré que le SNP 1 a la même information *a priori* que les autres SNPs, il n'est pas sélectionné car n'étant pas associé au trait considéré dans cette illustration. Avec la même information *a priori* de tous les SNPs, notre procédure parvient à sélectionner les SNPs associés au trait d'après le modèle étudié.
Si nous mettons une information *a priori* non conforme au modèle choisi ci-dessus, nous avons

SNP j	2	5	4	6	1	3
$1-p_j$	0.2	0.2	0.2	0.6	0.9	0.2
SNP sélectionné (oui/non)	oui	oui	oui	oui	oui	non

La procédure sélectionne dans cet ordre les SNPs 2, 5, 4, 6, et 1. Le SNP 1 qui n'est pas associé au trait d'intérêt est quand bien même sélectionné par la procédure du fait qu'il a une très forte probabilité *a priori* d'être causal $1-p_1 = 0.9$. Cependant le SNP 1 a été sélectionné qu'à la cinquième étape de la procédure de selection de SNP. Le SNP 5 qui a une plus faible probabilité *a priori* $1-p_5 = 0.2$ mais associé au trait grâce au modèle est sélectionné avant le SNP 1.

3.5 Étude d'impact de notre procédure de sélection de SNPs

Dans cette section, nous montrons à travers quelques scénarios l'impact de la correction de la loi asymptotique des statistiques d'association dans la sélection de SNPs associés à un trait d'intérêt. Pour cela nous appliquons notre procédure de sélection de SNPs avec la nouvelle loi asymptotique obtenue dans ce mémoire 2.7 et le résultat obtenu antérieurement par [28], [30], [27], [21].
Considérons l'ensemble des SNPs, $SNP = \{1,2,3,4,5,6\}$ dont la fréquence des allèles mineures est $f = \{f_1,f_2,f_3,f_4,f_5,f_6\}$ et la matrice de corrélation entre ces SNPs est notée V. Comme précedemment, la matrice de corrélation V est définie telle que la corrélation entre tous les SNPs est égale à 0.5 et la fréquence des allèles mineurs f des cinq SNPs est donnée par $f = (0.14, 0.39, 0.20, 0.44, 0.47, 0.02)$. Supposons que le trait Y est uniquement associé aux SNPs 1, 2 et 3.
Pour un nombre NB de répétition, nous calculons la proportion de fois que nous sélectionnons les 3 bon sSNPs avec notre procédure de sélection. Nous notons nb le pourcentage que la procédure sélectionne 3 SNPs en utilisant la loi asymptotique trouvée dans ce mémoire et nb_0 le pourcentage que la procédure sélectionne 3 SNPs en utilisant la loi asymptotique sans correction donnée par A.4, A.5 et A.6. Nous supposons aussi que les 6 SNPs ont la même probabilité *a priori* 0.5. Notre objectif est d'identifier les circonstances pour lesquelles l'utilisation de la bonne loi a un impact significatif dans la sélection de SNPs.
La procédure de sélection de SNPs va explorer les trois première étapes de sélection de SNPs. Nous présentons les lois asymptotiques utilisées dans les trois premières étapes de sélection pour calculer nb et nb_0 dans l'annexe A.4.

Exemple 1

Paramètres :
$NB = 1000$, $N = 1500$, $m = 6$, $\sigma_e = 1$, $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0.09$, $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$, $p_j = 0.5$ et $\sigma_{\lambda_j} = 0.1$, $j = 1, \cdots, m$.
Donc nos résultats de simulations dans ce cas montrent bien que nous avons une amélioration dans la sélection des trois SNPs significatifs associés au trait d'intérêt.

Nous donnons ici le pourcentage de sélection des SNPs 1, 2 et 3 supposés associés au trait.

SNP	nb	nb_0
(1,2,3)	85.2	78.7

En se trompant au moins une fois et au plus deux fois dans la sélection des SNPs 1, 2 et 3, nous obtenons les pourcentages de sélections suivantes

SNP	nb	nb_0	SNP	nb	nb_0
(1,2,4)	1.8	1.7	(1,4,5)	0	0
(1,2,5)	1.8	3.5	(1,4,6)	0	0
(1,2,6)	0.7	1.4	(1,5,6)	0	0
(1,3,4)	2.4	3.1	(2,4,5)	0	0
(1,3,5)	2.5	3.2	(2,4,6)	0	0.03
(1,3,6)	1.6	2.8	(2,5,6)	0	0
(2,3,4)	1	1.3	(3,4,5)	0	0
(2,3,5)	1.2	1.9	(3,4,6)	0.1	0.1
(2,3,6)	1.7	2.3	(3,5,6)	0	0

Enfin, si nous ne sélectionnons pas aucun des SNPs associés, nous avons le résultat suivant.

SNP	nb	nb_0
(4,5,6)	0	0

Nous remarquons qu'avec ou sans correction sur les paramètres de sélection de SNPs, la procédure trouve le plus souvent les SNPs associés au trait. Cependant avec notre correction, nous avons une augmentation de 6 points de pourcentage dans la sélection des SNPs 1, 2 et 3. Cette simulation illustre bien le gain acquis dans la sélection de SNPs causants.

Exemple 2

Dans cet exemple, nous étudions l'impact de la correction de la distribution de la statistique d'association pour différentes valeurs de l'effet des SNPs. Pour cela considérons le modèle de données et les mêmes paramètres de l'exemple 1. Ensuite nous fixons les paramètres $\beta_4 = \beta_5 = \beta_6 = 0$ et nous faisons varier les valeurs de β_0, β_1, β_2 et β_3. Nous obtenons le

tableau ci-dessous donnant nb et nb_0 pour les SNPs 1, 2 et 3.

$\beta_0 = \beta_1 = \beta_2 = \beta_3$	nb	nb_0
0.01	5.2	4.8
0.1	75.2	66.5
0.2	98.5	99.1
0.5	90.7	97.1

Notons que nous avons la même probabilité *a priori* 0.5 pour les 6 SNPs étudiés dans cet exemple. Pour de faibles effets du SNP, nous observons une légère amélioration dans la sélection des SNPs 1, 2 et 3. Pour des valeurs supérieures à 0.2, cet exemple montre que la correction peut être tr'es mauvais dans la sélection. Cependant avec la correction, la procédure sélectionne les SNPs 1, 2 et 3 à 90% du temps.
La correction sur la distribution de la statistique d'association peut apporter dans certains cas une amélioration dans la sélection des SNPs. Cependant nous n'avons pas une amélioration dans d'autres configurations pour l'effet du SNP et la probabilité *a priori*.

Conclusion et perspectives

Nous avons présenté dans ce travail une procédure itérative et conditionnelle qui permet de sélectionner un ensemble restreint de SNPs dans GWAS pour un trait d'intérêt donné. Une correction dans la loi asymptotique des statistiques d'association permet d'améliorer la recherche de SNP causal. La loi *a priori* choisie ici nous permet d'avoir des expressions explicites pour la loi *a posteriori*.

Pour illustrer notre procédure de sélection, nous utilisons une fonction qui génère aléatoirement des SNPs. Ensuite nous calculons un trait d'intérêt à partir de SNPs générés. Nos simulations montrent que notre procédure de sélection de SNPs peut aider à la priorisation des SNPs potentiellement causaux. Cependant, il faut noter que la correction de la loi asymptotique des statistiques d'association n'apporte pas d'amélioration dans la sélection de SNPs causants pour de fortes tailles d'effets du SNP.

Notre travail ne nous permet pas de déterminer le nombre K de SNPs qu'il faut piger dans les m SNPs de GWAS pour capturer tous les SNPs causaux. Plusieurs pistes de recherches peuvent venir compléter notre travail. Il serait intéressant de chercher une méthode pour déterminer K. Aussi élaborer un package R grâce à notre procédure de sélection pour aider à prioriser des SNPs est envisageable. Ensuite étudier une méthode de sélection de SNPs pour le cas des variants rares. Dans le cas des variants rares, l'effet d'un SNP causant est très faible avec une forte variabilité. De ce fait, il n'est pas intéressant d'utiliser une statistique de Wald dans la méthode de sélection de variants rares. Enfin, il sera utile de généraliser notre travail pour des données GWAS dépendantes.

Annexe A

Résultats

Dans notre travail, nous considérons que le trait Y est donné ainsi que les m SNPs.

A.1 Étape 1

Nous considérons les modèles d'hypothèses ci-dessous dont le trait est indépendant du SNP,

$$Y_i = \beta_{0j1} + \beta_{1j1} X_{ij1} + \varepsilon_{ij1}, \qquad j = 1, \cdots, m$$

où $\varepsilon_{ij1} \sim \mathcal{N}\left(0, \sigma^2_{j1}\right)$ et $\sigma_{j1} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape 1.

Soit $\tilde{X}_{j1} = \begin{pmatrix} 1 & X_{1j1} \\ \vdots & \vdots \\ 1 & X_{Nj1} \end{pmatrix}$ donc $\tilde{X}^T_{j1}\tilde{X}_{j1} = N \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

L'estimateur de l'effet du SNP j, $j = 1, \cdots, m$ est défini par

$$\begin{aligned} \hat{\beta}_{1j1} &= (0,1)\left(\tilde{X}^T_{j1}\tilde{X}_{j1}\right)^{-1}\tilde{X}^T_{j1}Y \\ &= N^{-1}\sum_{i=1}^{N} X_{ij1}Y_i \quad \text{car} \quad \tilde{X}^T_{j1}\tilde{X}_{j1} = N\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= N^{-1}X^T_{j1}Y \quad \text{avec} \quad X^T_{j1} = (X_{1j1}, \cdots, X_{Nj1}). \end{aligned}$$

Donc nous avons

$$\hat{\beta}_{1j1} \sim \mathcal{N}\left(\beta_{1j1}, \sigma^2_{j1}/N\right). \tag{A.1}$$

En effet,

$$\begin{aligned} \mathbb{E}\left(\hat{\beta}_{1j1}\right) &= X^T_{j1}\mathbb{E}(Y)/N \\ &= \beta_{1j1}\sum_{i=1}^{N} X^2_{ij1}/N \\ &= \beta_{1j1} \quad \text{car} \quad \sum_{i=1}^{N} X^2_{ij1} = N \end{aligned}$$

et

$$\begin{aligned} Var\left(\hat{\beta}_{1j1}\right) &= X_{j1}^T Var\left(Y\right) X_{j1}/N^2 \\ &= \sum_{i=1}^{N} X_{ij1}^2 \sigma_{j1}^2/N^2 \\ &= \sigma_{j1}^2/N. \end{aligned}$$

D'où

$$\hat{\beta}_{1j1} \sim \mathcal{N}\left(\beta_{1j1}, \frac{\sigma_{j1}^2}{N}\right).$$

La variance des erreurs est estimée par $\hat{\sigma}_{j1}^2 = \frac{1}{N-2}\left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)^T \left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)$ avec $\hat{\beta}_{j1}^T = (\hat{\beta}_{0j1}, \hat{\beta}_{1j1})$.
Nous avons $\mathbb{E}\left(\hat{\sigma}_{j1}^2\right) = \sigma_{j1}^2$ et $Var\left(\hat{\sigma}_{j1}^2\right) = \frac{2\sigma_{j1}^4}{N-2}$.
La statistique d'association de Wald

$$\hat{Z}_{j1} = \hat{\beta}_{1j1}/\widehat{SE}\left(\hat{\beta}_{1j1}\right)$$

se réécrit comme suit

$$\hat{Z}_{j1} = \frac{\hat{\beta}_{1j1}}{\sqrt{\hat{\sigma}_{j1}^2/N}}.$$

Posons

$$\hat{Z}_{j1} = g\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right)$$

où g est une fonction à deux variables définie par

$$g\left(a, b\right) = \frac{a}{\sqrt{b/N}}.$$

Par un développement de Taylor d'ordre 1 sur $\left(\beta_{1j1}, \sigma_{j1}^2\right)$, nous obtenons

$$g\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) \simeq g\left(\beta_{1j1}, \sigma_{j1}^2\right) + \left(\hat{\beta}_{1j1} - \beta_{1j1}\right) \frac{\partial g}{\partial \hat{\beta}_{1j1}}\left(\beta_{1j1}, \sigma_{j1}^2\right) + \left(\hat{\sigma}_{j1}^2 - \sigma_{j1}^2\right) \frac{\partial g}{\partial \hat{\sigma}_{j1}^2}\left(\beta_{1j1}, \sigma_{j1}^2\right).$$

Alors la moyenne asymptotique de $\hat{Z}_{j1}$ est

$$\begin{aligned} \mathbb{E}\left(\hat{Z}_{j1}\right) &= g\left(\beta_{1j1}, \sigma_{j1}^2\right) \\ &= \sqrt{N}\lambda_{j1} \end{aligned}$$

avec $\lambda_{j1} = \beta_{1j1}/\sigma_{j1}$.
Montrons que $Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) = 0$.
Nous avons

$$\begin{aligned} Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) &= Cov\left(\hat{\beta}_{1j1}, \frac{1}{N-2}\left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)^T \left(Y - \tilde{X}_{j1}\hat{\beta}_{j1}\right)\right) \quad \text{avec} \quad \hat{\beta}_{j1}^T = \left(\hat{\beta}_{0j1}, \hat{\beta}_{1j1}\right) \\ Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) &= \frac{1}{N-2} Cov\left(\hat{\beta}_{1j1}, Y^T Y - 2\hat{\beta}_{j1}^T \tilde{X}_{j1}^T Y + \hat{\beta}_{j1}^T \tilde{X}_{j1}^T \tilde{X}_{j1} \hat{\beta}_{j1}\right) \\ (N-2) Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) &= Cov\left(\hat{\beta}_{1j1}, Y^T Y\right) - 2Cov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \tilde{X}_{j1}^T Y\right) + Cov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \tilde{X}_{j1}^T \tilde{X}_{j1} \hat{\beta}_{j1}\right). \end{aligned}$$

Comme $\hat{\beta}_{j1} = \left(\tilde{X}_{j1}^T \tilde{X}_{j1}\right)^{-1} \tilde{X}_{j1}^T Y$ et $\tilde{X}_{j1}^T \tilde{X}_{j1} = N \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, alors $\tilde{X}_{j1}^T Y = N\hat{\beta}_{j1}$.

Donc

$$\begin{aligned} (N-2)Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) &= Cov\left(\hat{\beta}_{1j1}, Y^T Y\right) - 2NCov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \hat{\beta}_{j1}\right) + NCov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \hat{\beta}_{j1}\right) \\ &= Cov\left(\hat{\beta}_{1j1}, Y^T Y\right) - NCov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \hat{\beta}_{j1}\right). \end{aligned}$$

Or $\hat{\beta}_{j1}^T \hat{\beta}_{j1} = Y^T \tilde{X}_{j1} \tilde{X}_{j1}^T Y / N^2 = Y^T Y / N$, alors

$$\begin{aligned} (N-2)Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) &= Cov\left(\hat{\beta}_{1j1}, Y^T Y\right) - 2NCov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \hat{\beta}_{j1}\right) + NCov\left(\hat{\beta}_{1j1}, \hat{\beta}_{j1}^T \hat{\beta}_{j1}\right) \\ &= Cov\left(\hat{\beta}_{1j1}, Y^T Y\right) - NCov\left(\hat{\beta}_{1j1}, Y^T Y / N\right) \\ &= 0. \end{aligned}$$

D'où $Cov\left(\hat{\beta}_{1j1}, \hat{\sigma}_{j1}^2\right) = 0$.

Par la suite, la variance asymptotique de $\hat{Z}_{j1}$ est donnée par

$$Var\left(\hat{Z}_{j1}\right) = \left[\frac{\partial g}{\partial \hat{\beta}_{1j1}}\left(\beta_{1j1}, \sigma_{j1}^2\right)\right]^2 Var\left(\hat{\beta}_{1j1}\right) + \left[\frac{\partial g}{\partial \hat{\sigma}_{j1}^2}\left(\beta_{1j1}, \sigma_{j1}^2\right)\right]^2 Var\left(\hat{\sigma}_{j1}^2\right)$$

or

$$\frac{\partial g}{\partial \beta_{1j1}}\left(\beta_{1j1}, \sigma_{j1}^2\right) = \frac{1}{\sqrt{\sigma_{j1}^2 / N}}$$

et

$$\frac{\partial g}{\partial \sigma_{j1}^2}\left(\beta_{1j1}, \sigma_{j1}^2\right) = -\frac{\sqrt{N}\beta_{1j1}}{2\left(\sigma_{j1}^2\right)^{3/2}}.$$

Donc

$$\begin{aligned} Var\left(\hat{Z}_{j1}\right) &= 1 + \frac{2N\beta_{1j1}^2 \sigma^4}{4\sigma^6} \\ &= 1 + \frac{1}{2}\lambda_{j1}^2. \end{aligned}$$

D'où

$$\hat{Z}_{j1} \sim \mathcal{N}\left(\sqrt{N}\lambda_{j1}, 1 + \frac{1}{2}\lambda_{j1}^2\right)$$

A.2 Étape 2

Nous supposons que le SNP 1 est choisi à l'étape 1 et son code est noté par X_{i12}. Considérons les modèles d'hypothèses ci-dessous dont le trait est indépendant du SNP,

$$Y_i = \beta_{0j2} + \beta_{1j2} X_{i12} + \beta_{2j2} X_{ij2} + \varepsilon_{ij2}, \qquad j = 2, \cdots, m$$

où $\varepsilon_{ij2} \sim \mathcal{N}\left(0,\sigma_{j2}^2\right)$, $\sigma_{j2} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape 2 et $\forall i = 1,\cdots,N$.

Soit $\tilde{X}_{j2} = \begin{pmatrix} 1 & X_{112} & X_{1j2} \\ \vdots & \vdots & \\ 1 & X_{N12} & X_{Nj2} \end{pmatrix}$ donc $\tilde{X}_{j2}^T\tilde{X}_{j2} = N\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V_{j2} \end{pmatrix}$

avec $V_{j2} = \begin{pmatrix} 1 & r_{1j2} \\ r_{1j2} & 1 \end{pmatrix}$, $\mathbf{0} = (0,0)^T$ et $r_{lq2} = \frac{1}{N}\sum_{i=1}^{N} X_{il2}X_{iq2}$, $l,q \in \{1,j\}$, $j = 2,\cdots,m$.

L'effet du SNP j et son estimateur est défini par $\beta_{j2} = (\beta_{0j2},\beta_{1j2},\beta_{2j2})^T$ et $\hat{\beta}_{j2} = \left(\tilde{X}_{j2}^T\tilde{X}_{j2}\right)^{-1}\tilde{X}_{j2}^T Y$.

Comme la variance des erreurs du modèle est estimée par

$$\hat{\sigma}_{j2}^2 = \frac{\left(Y - \tilde{X}_{j2}\hat{\beta}_{j2}\right)^T\left(Y - \tilde{X}_{j2}\hat{\beta}_{j2}\right)}{N-3}$$

alors $\hat{\beta}_{j2} \sim \mathcal{N}\left(\beta_{j2}, \sigma_{j2}^2\left(\tilde{X}_{j2}^T\tilde{X}_{j2}\right)^{-1}\right)$ avec $\left(\tilde{X}_{j2}^T\tilde{X}_{j2}\right)^{-1} = N^{-1}\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & W_{j2} \end{pmatrix}$ et $W_{j2} = V_{j2}^{-1}$,

$V_{j2}^{-1} = \frac{1}{1-r_{1j2}^2}\begin{pmatrix} 1 & -r_{1j2} \\ -r_{1j2} & 1 \end{pmatrix}$.

Soit W_{j2lq} l'élément à la position (l,q) de la matrice W_{j2}. Pour $(l,q) \in \{1,2\}^2$, $Var\left(\hat{\beta}_{lj2}\right) = \frac{1}{N}\sigma_{j2}^2 W_{j2ll}$ alors $Var\left(\hat{\beta}_{1j2}\right) = Var\left(\hat{\beta}_{2j2}\right) = \frac{1}{N}\sigma_{j2}^2\frac{1}{1-r_{1j2}^2}$. De plus $Cov\left(\hat{\beta}_{lj2},\hat{\beta}_{qj2}\right) = \frac{1}{N}\sigma_{j2}^2 W_{j2lq}$ donc $Cov\left(\hat{\beta}_{1j2},\hat{\beta}_{2j2}\right) = \frac{1}{N}\sigma_{j2}^2\frac{-r_{1j2}^2}{1-r_{1j2}^2}$. En plus, nous avons $(N-3)\frac{\hat{\sigma}_{j2}^2}{\sigma_{j2}^2} \sim \chi_{N-3}^2$ donc $Var\left(\hat{\sigma}_{j2}^2\right) \simeq \frac{2\sigma_{j2}^4}{N}$ et $Cov\left(\hat{\beta}_{lj2},\hat{\sigma}_{j2}^2\right) = 0$.

La statistique de test peut être écrite comme suit

$$\hat{Z}_{lj2} = \frac{\sqrt{N}}{\sqrt{W_{j2ll}}}\frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}_{j2}^2}} \qquad l = 1,2.$$

Nous pouvons calculer la moyenne de la statistique d'association pour le SNP l, $l = 1,2$ par

$$\begin{aligned} \mathbb{E}\left(\hat{Z}_{lj2}\right) &= \frac{\sqrt{N}}{\sqrt{W_{j2ll}}}\mathbb{E}\left(\frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}_{j2}^2}}\right) \\ &= \frac{\sqrt{N}}{\sqrt{W_{j2ll}}}\mathbb{E}\left(g\left(\hat{\beta}_{lj2},\hat{\sigma}_{j2}^2\right)\right) \end{aligned}$$

avec $g(a,b) = \frac{a}{\sqrt{b}}$.

Par un développement de Taylor d'ordre 1 en (a,b), nous obtenons asymptotiquement

$$g(\hat{a},\hat{b}) \simeq g(a,b) + \frac{\partial g}{\partial a}(a,b)\,(\hat{a}-a) + \frac{\partial g}{\partial b}(a,b)\left(\hat{b}-b\right)$$

donc

$$\mathbb{E}\left(g\left(\hat{\beta}_{lj2},\hat{\sigma}_{j2}^2\right)\right) = \frac{\beta_{lj2}}{\sqrt{\sigma_{j2}^2}}$$

et

$$\begin{aligned}\mathbb{E}\left(\hat{Z}_{lj2}\right) &= \frac{\sqrt{N}}{\sqrt{W_{j2ll}}}\frac{\beta_{lj2}}{\sqrt{\sigma^2_{j2}}} \\ &= \sqrt{N}\sqrt{1-r^2_{lj2}}\lambda_{lj2}.\end{aligned}$$

Nous calculons aussi la variance de la statistique de test par

$$\begin{aligned}Var\left(\hat{Z}_{lj2}\right) &= \frac{N}{W_{j2ll}}Var\left(\frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}^2_{j2}}}\right) \\ &= \frac{N}{W_{j2ll}}Var\left(g\left(\hat{\beta}_{lj2},\hat{\sigma}^2_{j2}\right)\right).\end{aligned}$$

En utilisant le même principe de calcul pour montrer que $Cov\left(\hat{\beta}_{1j1},\hat{\sigma}^2_{j1}\right) = 0$, nous avons aussi $Cov\left(\hat{a},\hat{b}\right) = 0$, alors

$$Var\left(g\left(\hat{a},\hat{b}\right)\right) = \left[\frac{\partial g}{\partial a}(a,b)\right]^2 Var\left(\hat{a}\right) + \left[\frac{\partial g}{\partial b}(a,b)\right]^2 Var\left(\hat{b}\right).$$

Nous avons

$$\frac{\partial g}{\partial a}(a,b) = \frac{1}{\sqrt{b}} \qquad \frac{\partial g}{\partial b}(a,b) = -\frac{1}{2}\frac{a}{b\sqrt{b}}$$

et

$$Var\left(g\left(\hat{\beta}_{lj2},\hat{\sigma}^2_{j2}\right)\right) = \frac{1}{\sigma^2_{j2}}\frac{1}{N}\sigma^2_{j2}W_{j2ll} + \frac{1}{2}\frac{1}{N}\frac{\beta^2_{lj2}}{\left(\hat{\sigma}^2_{j2}\right)^3}\sigma^4_{j2}$$

donc

$$\begin{aligned}Var\left(\hat{Z}_{lj2}\right) &= 1 + \frac{1}{2}\frac{\lambda^2_{l2}}{W_{j2ll}} \\ &= 1 + \frac{1}{2}\left(1-r^2_{1j2}\right)\lambda^2_{lj2}.\end{aligned}$$

La covariance entre $\hat{Z}_{lj2}$ et $\hat{Z}_{qj2}$ est donnée par

$$Cov\left(\hat{Z}_{lj2},\hat{Z}_{qj2}\right) = \frac{N}{\sqrt{W_{j2ll}}\sqrt{W_{j2qq}}}Cov\left(\frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}^2_{j2}}},\frac{\hat{\beta}_{qj2}}{\sqrt{\hat{\sigma}^2_{j2}}}\right)$$

De même, nous avons

$$Cov\left(g(\hat{a}_l,\hat{b}),g(\hat{a}_q,\hat{b})\right) = \frac{\partial g}{\partial a_l}(a_l,b)\frac{\partial g}{\partial a_q}(a_q,b)Cov\left(\hat{a}_l,\hat{a}_l\right) + \frac{\partial g}{\partial b}(a_l,b)\frac{\partial g}{\partial b}(a_q,b)Var\left(\hat{b}\right)$$

comme

$$Cov\left(\hat{\beta}_{lj2},\hat{\beta}_{qj2}\right) = \frac{1}{N}W_{j2lq}\sigma^2_{j2}$$

alors

$$\begin{aligned} Cov\left(\frac{\hat{\beta}_{lj2}}{\sqrt{\hat{\sigma}^2_{j2}}},\frac{\hat{\beta}_{qj2}}{\sqrt{\hat{\sigma}^2_{j2}}}\right) &= \frac{1}{\sqrt{\sigma^2_{j2}}}\frac{1}{\sqrt{\sigma^2_{j2}}}\frac{1}{N}W_{j2lq}\sigma^2_{j2} + \left(-\frac{1}{2}\frac{\beta_{lj2}}{\sigma^2_{j2}\sqrt{\sigma^2_{j2}}}\right)\left(-\frac{1}{2}\frac{\beta_{qj2}}{\sigma^2_{j2}\sqrt{\sigma^2_{j2}}}\right)\frac{2\sigma^4_{j2}}{N} \\ &= \frac{1}{N}W_{j2lq} + \frac{1}{2N}\lambda_{lj2}\lambda_{qj2} \end{aligned}$$

donc

$$Cov\left(\hat{Z}_{lj2},\hat{Z}_{qj2}\right) = \frac{W_{j2lq}}{\sqrt{W_{j2ll}}\sqrt{W_{j2qq}}} + \frac{1}{2}\frac{\lambda_{lj2}\lambda_{qj2}}{\sqrt{W_{j2ll}}\sqrt{W_{j2qq}}}$$

d'où

$$\begin{aligned} Cov\left(\hat{Z}_{1j2},\hat{Z}_{2j2}\right) &= \frac{\frac{-r_{1j2}}{1-r^2_{1j2}}}{\sqrt{\frac{1}{1-r^2_{1j2}}}\sqrt{\frac{1}{1-r^2_{1j2}}}} + \frac{1}{2}\frac{\lambda_{1j2}\lambda_{2j2}}{\sqrt{\frac{1}{1-r^2_{1j2}}}\sqrt{\frac{1}{1-r^2_{1j2}}}} \\ &= -r_{1j2} + \frac{1}{2}\left(1-r^2_{1j2}\right)\lambda_{1j2}\lambda_{2j2}. \end{aligned}$$

Considérons le vecteur des statistiques d'association noté par $\hat{Z}_{j2} = \left(\hat{Z}_{1j2},\hat{Z}_{2j2}\right)^T$. Donc

$$\hat{Z}_{j2} = \sqrt{N}D_{j2}\hat{\lambda}_{j2}$$

où $\hat{\lambda}_{j2} = \left(\hat{\lambda}_{1j2},\hat{\lambda}_{2j2}\right)^T$ et

$$D_{j2} = \begin{pmatrix} \sqrt{1-r^2_{1j2}} & 0 \\ 0 & \sqrt{1-r^2_{1j2}} \end{pmatrix}$$

Comme précédemment, nous montrons que

$$\hat{Z}_{j2} \sim \mathcal{N}_2\left(M_{j2},\Sigma_{j2}\right) \tag{A.2}$$

avec $M_{j2} = \sqrt{N}\sqrt{1-r^2_{1j2}}\lambda_{j2}$,

$$\begin{aligned} \Sigma_{j2} &= \begin{pmatrix} 1+\frac{1}{2}\left(1-r^2_{1j2}\right)\lambda^2_{1j2} & -r_{1j2}+\frac{1}{2}\left(1-r^2_{1j2}\right)\lambda_{1j2}\lambda_{2j2} \\ -r_{1j2}+\frac{1}{2}\left(1-r^2_{1j2}\right)\lambda_{1j2}\lambda_{2j2} & 1+\frac{1}{2}\left(1-r^2_{1j2}\right)\lambda^2_{2j2} \end{pmatrix} \\ &= \begin{pmatrix} 1 & -r_{1j2} \\ -r_{1j2} & 1 \end{pmatrix} + \frac{1}{2}\left(1-r^2_{1j2}\right)\lambda^T_{j2}\lambda_{j2} \end{aligned}$$

où $\lambda_{j2} = (\lambda_{1j2},\lambda_{2j2})^T$.

A.3 Étape k

Nous supposons que les SNPs 1,2,$\cdots$,$k-1$ ont été choisi au $k-1$ premières étapes et leurs codes sont notés par $X_{i1k}, X_{i2k}, \cdots, X_{i,k-1,k}$. Considérons les modèles d'hypothèses ci-dessous dont le trait est indépendant du SNP,

$$Y_i = \beta_{0jk} + \beta_{1jk}X_{i1k} + \cdots + \beta_{k-1,j,k}X_{i,k-1,k} + \beta_{kjk}X_{ijk} + \varepsilon_{ijk}, \qquad j = k,k+1,\cdots,m$$

où $\varepsilon_{ijk} \sim \mathcal{N}\left(0,\sigma^2_{jk}\right)$, $\sigma_{jk} > 0$ est l'écart type des erreurs pour le j-ème modèle à l'étape k et $\forall i = 1,\cdots,N$.

Soit $\tilde{X}_{jk} = \begin{pmatrix} 1 & X_{11k} & \cdots & X_{1,k-1,k} & X_{1jk} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & X_{N1k} & \cdots & X_{N,k-1,k} & X_{Njk} \end{pmatrix}$ donc $\tilde{X}^T_{jk}\tilde{X}_{jk} = N \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V_{jk} \end{pmatrix}$

avec $V_{jk} = \begin{pmatrix} 1 & r_{12k} & \cdots & r_{1,k-1,k} & r_{1jk} \\ r_{12k} & \ddots & & & \\ \vdots & & \ddots & & \\ r_{1,k-1,k} & & & \ddots & \\ r_{1j2} & & & & 1 \end{pmatrix}$, $\mathbf{0} = (0,\cdots,0)^T$ est le vecteur nul de longueur k et $r_{lqk} = \frac{1}{N}\sum_{i=1}^{N} X_{ilk}X_{iqk}$, $(l,q) \in \{1,2,\cdots,k-1,j\}^2$.

L'effet du SNP j et son estimateur est défini par $\beta_{jk} = (\beta_{0jk},\beta_{1jk},\cdots,\beta_{kjk})$ et $\hat{\beta}_{jk} = \left(\tilde{X}^T_{jk}\tilde{X}_{jk}\right)^{-1}\tilde{X}^T_{jk}Y$. Comme la variance des erreurs du modèle est estimée par

$$\hat{\sigma}^2_{jk} = \frac{\left(Y - \tilde{X}_{jk}\hat{\beta}_{jk}\right)^T\left(Y - \tilde{X}_{jk}\hat{\beta}_{jk}\right)}{N-(k+1)}$$

alors $\hat{\beta}_{jk} \sim \mathcal{N}\left(\beta_{jk}, \sigma^2_{jk}\left(\tilde{X}^T_{jk}\tilde{X}_{jk}\right)^{-1}\right)$ avec $\left(\tilde{X}^T_{jk}\tilde{X}_{jk}\right)^{-1} = N^{-1}\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & W_{jk} \end{pmatrix}$ et $W_{jk} = V^{-1}_{jk}$.

Soit W_{j2lq} l'élément à la position (l,q) de la matrice W_{jk}. Pour $(l,q) \in \{1,2,\cdots,k\}^2$, $Var\left(\hat{\beta}_{ljk}\right) = \frac{1}{N}\sigma^2_{jk}W_{jkll}$ et $Cov\left(\hat{\beta}_{ljk},\hat{\beta}_{qjk}\right) = \frac{1}{N}\sigma^2_{j2}W_{j2lq}$. En plus, nous avons $(N-k-1)\frac{\hat{\sigma}^2_{jk}}{\sigma^2_{jk}} \sim \chi^2_{N-k-1}$ donc $Var\left(\hat{\sigma}^2_{jk}\right) \simeq \frac{2\sigma^4_{jk}}{N}$ et $Cov\left(\hat{\beta}_{ljk},\hat{\sigma}^2_{jk}\right) = 0$.

La statistique de test peut être écrite comme suit

$$\hat{Z}_{ljk} = \frac{\sqrt{N}}{\sqrt{W_{jkll}}}\frac{\hat{\beta}_{ljk}}{\sqrt{\hat{\sigma}^2_{jk}}} \qquad l = 1,2,\cdots,k.$$

Nous pouvons calculer la moyenne de la statistique de test pour le SNP l, $l = 1,2,\cdots,k$ par

$$\begin{aligned} \mathbb{E}\left(\hat{Z}_{ljk}\right) &= \frac{\sqrt{N}}{\sqrt{W_{jkll}}}\mathbb{E}\left(\frac{\hat{\beta}_{ljk}}{\sqrt{\hat{\sigma}^2_{jk}}}\right) \\ &= \frac{\sqrt{N}}{\sqrt{W_{jkll}}}\mathbb{E}\left(g\left(\hat{\beta}_{ljk},\hat{\sigma}^2_{jk}\right)\right) \end{aligned}$$

avec $g(a,b) = \frac{a}{\sqrt{b}}$.

Par un développement de Taylor d'ordre 1 en (a,b), nous obtenons asymptotiquement

$$g(\hat{a},\hat{b}) = g(a,b) + \frac{\partial g}{\partial a}(a,b)\left(\hat{a}-a\right) + \frac{\partial g}{\partial b}(a,b)\left(\hat{b}-b\right)$$

donc

$$\mathbb{E}\left(g\left(\hat{\beta}_{ljk},\hat{\sigma}^2_{jk}\right)\right) = \frac{\beta_{ljk}}{\sqrt{\sigma^2_{jk}}}$$

et

$$\begin{aligned}\mathbb{E}\left(\hat{Z}_{ljk}\right) &= \frac{\sqrt{N}}{\sqrt{W_{jkll}}}\frac{\beta_{ljk}}{\sqrt{\sigma^2_{jk}}} \\ &= \frac{\sqrt{N}}{\sqrt{W_{jkll}}}\lambda_{ljk}.\end{aligned}$$

Nous calculons aussi la variance de la statistique de test par

$$\begin{aligned}Var\left(\hat{Z}_{ljk}\right) &= \frac{N}{W_{jkll}}Var\left(\frac{\hat{\beta}_{ljk}}{\sqrt{\hat{\sigma}^2_{jk}}}\right) \\ &= \frac{N}{W_{jkll}}Var\left(g\left(\hat{\beta}_{ljk},\hat{\sigma}^2_{jk}\right)\right)\end{aligned}$$

Comme montrer dans l'étape 1 que $Cov\left(\hat{\beta}_{1j1},\hat{\sigma}^2_{j1}\right)=0$, nous avons aussi $Cov\left(\hat{a},\hat{b}\right)=0$, alors

$$Var\left(g\left(\hat{a},\hat{b}\right)\right)=\left[\frac{\partial g}{\partial a}(a,b)\right]^2 Var\left(\hat{a}\right)+\left[\frac{\partial g}{\partial b}(a,b)\right]^2 Var\left(\hat{b}\right).$$

Nous avons

$$\frac{\partial g}{\partial a}(a,b)=\frac{1}{\sqrt{b}} \qquad \frac{\partial g}{\partial b}(a,b)=-\frac{1}{2}\frac{a}{b\sqrt{b}}$$

et

$$Var\left(g\left(\hat{\beta}_{ljk},\hat{\sigma}^2_{jk}\right)\right)=\frac{1}{\sigma^2_{jk}}\frac{1}{N}\sigma^2_{jk}W_{jkll}+\frac{1}{2}\frac{1}{N}\frac{\beta^2_{ljk}}{\left(\hat{\sigma}^2_{jk}\right)^3}\sigma^4_{jk}$$

donc

$$Var\left(\hat{Z}_{ljk}\right)=1+\frac{1}{2}\frac{\lambda^2_{ljk}}{W_{jkll}}$$

De même, nous montrons que

$$Cov\left(\hat{Z}_{ljk},\hat{Z}_{qjk}\right)=\frac{W_{jklq}}{\sqrt{W_{jkll}}\sqrt{W_{jkqq}}}+\frac{1}{2}\frac{\lambda_{ljk}\lambda_{qjk}}{\sqrt{W_{jkll}}\sqrt{W_{jkqq}}}$$

Considérons le vecteur des statistiques d'associations noté par $\hat{Z}_{jk}=\left(\hat{Z}_{1jk},\hat{Z}_{2jk},\cdots,\hat{Z}_{kjk}\right)^T$. Donc

$$\hat{Z}_{jk}=\sqrt{N}D_{jk}\hat{\lambda}_{jk}$$

où $D_{jk}=Diag\left(d_{jk}\right)$ est la matrice diagonale dont la diagonale principale est le vecteur d_{jk}, $d_{jk}=\left(W^{-1/2}_{jk11},\cdots,W^{-1/2}_{jkkk}\right)^T$ et $\hat{\lambda}_{jk}=\left(\hat{\lambda}_{1jk},\hat{\lambda}_{2jk},\cdots,\hat{\lambda}_{kjk}\right)^T$. Comme précédemment, nous montrons que

$$\hat{Z}_{jk}\sim\mathcal{N}_k\left(M_{jk},\Sigma_{jk}\right) \tag{A.3}$$

avec $M_{jk} = D_{jk}\lambda_{jk}$ où D_{jk} est une matrice diagonale de dimensions $k \times k$,

$$D_{jk} = \begin{pmatrix} \frac{\sqrt{N}}{\sqrt{W_{jk11}}} & 0 & \cdots & 0 \\ 0 & \frac{\sqrt{N}}{\sqrt{W_{jk22}}} & & \vdots \\ \vdots & & \ddots & \\ 0 & & & \frac{\sqrt{N}}{\sqrt{W_{jkkk}}} \end{pmatrix}$$

et

$$\Sigma_{jk} = R^1_{jk} + \frac{1}{2} R^2_{jk} \lambda^T_{jk} \lambda_{jk}$$

avec R^1_{jk} et R^2_{jk} sont des matrices de dimensions $k \times k$ telles que

$$R^1_{jk} = \begin{pmatrix} 1 & \frac{W_{jk12}}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \cdots & \frac{W_{jk1k}}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} \\ \frac{W_{jk12}}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & 1 & & \\ \vdots & & \ddots & \\ \frac{W_{jk1k}}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} & & & 1 \end{pmatrix},$$

$$R^2_{jk} = \begin{pmatrix} \frac{1}{W_{jk11}} & \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \cdots & \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} \\ \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jk22}}} & \frac{1}{W_{jk11}} & & \\ \vdots & & \ddots & \\ \frac{1}{\sqrt{W_{jk11}}\sqrt{W_{jkkk}}} & & & \frac{1}{W_{jkkk}} \end{pmatrix}.$$

A.4 Loi asymptotique de la statistique d'association

Nous présentons ci-dessous les lois asymptotiques de la statistique d'association pour les trois premières étapes de la procédure de sélection de SNPs.

— Étape 1 :

Pour calculer nb, la procédure de sélection de SNPs utilise la loi asymptotique suivante

$$\hat{Z}_{j1} \sim \mathcal{N}\left(\sqrt{N}\lambda_{j1}, 1 + \frac{1}{2}\lambda^2_{j1}\right), \qquad j = 1, \cdots, m.$$

Pour calculer nb_0, la procédure de sélection de SNPs utilise à l'étape 1 la loi asymptotique suivante

$$\hat{Z}_{j1} \sim \mathcal{N}\left(\sqrt{N}\lambda_{j1}, 1\right), \qquad j = 1, \cdots, m. \tag{A.4}$$

— Étape 2 :

Pour calculer nb, la procédure de sélection de SNPs utilise la loi asymptotique suivante

$$\hat{Z}_{lj2} \sim \mathcal{N}\left(\sqrt{N}\sqrt{1 - r^2_{1j2}}\lambda_{lj2}, 1 + \frac{1}{2}\left(1 - r^2_{1j2}\right)\lambda^2_{lj2}\right), \qquad l = 1,2, \qquad j = 2, \cdots, m$$

avec r_{1j2} la corrélation entre le SNP 1 sélectionné à l'étape 1 et le SNP j, $j = 2, \cdots, m$.
De plus pour $j = 2, \cdots, m$ nous avons

$$\hat{Z}_{j2} \sim \mathcal{N}\left(\begin{pmatrix}\sqrt{N}\sqrt{1-r_{1j2}^2}\lambda_{1j2}\\ \sqrt{N}\sqrt{1-r_{1j2}^2}\lambda_{2j2}\end{pmatrix}, \begin{pmatrix}1+\frac{1}{2}\left(1-r_{1j2}^2\right)\lambda_{1j2}^2 & -r_{1j2}+\frac{1}{2}\left(1-r_{1j2}^2\right)\lambda_{1j2}\lambda_{2j2}\\ -r_{1j2}+\frac{1}{2}\left(1-r_{1j2}^2\right)\lambda_{1j2}\lambda_{2j2} & 1+\frac{1}{2}\left(1-r_{1j2}^2\right)\lambda_{2j2}^2\end{pmatrix}\right).$$

Pour calculer nb_0, la procédure de sélection de SNPs utilise à l'étape 2 la loi asymptotique suivante

$$\hat{Z}_{lj2} \sim \mathcal{N}\left(\sqrt{N}\sqrt{1-r_{1j2}^2}\lambda_{lj2}, 1\right) \qquad l = 1,2, \qquad j = 2, \cdots, m$$

et

$$\hat{Z}_{j2} \sim \mathcal{N}\left(\begin{pmatrix}\sqrt{N}\sqrt{1-r_{1j2}^2}\lambda_{1j2}\\ \sqrt{N}\sqrt{1-r_{1j2}^2}\lambda_{2j2}\end{pmatrix}, \begin{pmatrix}1 & r_{1j2}\\ r_{1j2} & 1\end{pmatrix}\right), \qquad j = 2, \cdots, m. \tag{A.5}$$

— Étape 3 :
Nous avons 1 et 2 les SNPs sélectionnés aux étapes 1 et 2 respectivement. $r = (r_{123}, r_{1j3}, r_{2j3})$, r_{123} la corrélation entre les SNPs 1 et 2, r_{1j3} la corrélation entre les SNPs 1 et j, r_{2j3} la corrélation entre les SNPs 2 et j. Soit

$$\begin{aligned}\alpha(r) &= det\left(\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & r_{123} & r_{1j3}\\ 0 & r_{123} & 1 & r_{2j3}\\ 0 & r_{1j3} & r_{2j3} & 1\end{pmatrix}\right)\\ &= \left(1-r_{2j3}^2\right) - r_{123}\left(r_{123} - r_{1j3}r_{2j3}\right) - r_{1j3}\left(r_{1j3} - r_{123}r_{2j3}\right)\end{aligned}$$

et $c_1(r) = \frac{1-r_{2j3}^2}{\alpha(r)}$, $c_2(r) = \frac{1-r_{1j3}^2}{\alpha(r)}$, $c_3(r) = \frac{1-r_{123}^2}{\alpha(r)}$.
Pour calculer nb, la procédure de sélection de SNPs utilise la loi asymptotique suivante

$$\hat{Z}_{lj3} \sim \mathcal{N}\left(\sqrt{N}\sqrt{c_l^{-1}(r)}\lambda_{lj3}, 1 + \frac{1}{2}c_l^{-1}(r)\lambda_{lj3}\right) \qquad l = 1,2,3 \qquad j = 3, \cdots, m.$$

De plus, $\hat{Z}_{j3} = \left(\hat{Z}_{1j3}, \hat{Z}_{2j3}, \hat{Z}_{3j3}\right)$ a une loi asymptotique normale multivariée. Les covariances entre les statistiques d'associations sont données par

$$\begin{aligned}Cov\left(\hat{Z}_{1j3}, \hat{Z}_{2j3}\right) &= \frac{r_{1j3}r_{2j3} - r_{123}}{c_1(r)c_2(r)\alpha(r)} + \frac{1}{2}\sqrt{c_1^{-1}(r)c_2^{-1}(r)}\lambda_{1j3}\lambda_{2j3}\\ Cov\left(\hat{Z}_{1j3}, \hat{Z}_{3j3}\right) &= \frac{r_{123}r_{2j3} - r_{1j3}}{c_1(r)c_3(r)\alpha(r)} + \frac{1}{2}\sqrt{c_1^{-1}(r)c_3^{-1}(r)}\lambda_{1j3}\lambda_{3j3}\\ Cov\left(\hat{Z}_{2j3}, \hat{Z}_{3j3}\right) &= \frac{r_{123}r_{2j3} - r_{2j3}}{c_2(r)c_3(r)\alpha(r)} + \frac{1}{2}\sqrt{c_2^{-1}(r)c_3^{-1}(r)}\lambda_{2j3}\lambda_{3j3}.\end{aligned}$$

Pour calculer nb_0, la procédure de sélection de SNPs utilise à l'étape 3 la loi asymptotique suivante

$$\hat{Z}_{j3} \sim \mathcal{N}\left(\begin{pmatrix}\sqrt{N}\sqrt{c_1^{-1}(r)}\lambda_{1j3}\\ \sqrt{N}\sqrt{c_2^{-1}(r)}\lambda_{2j3}\\ \sqrt{N}\sqrt{c_3^{-1}(r)}\lambda_{3j3}\end{pmatrix}, \begin{pmatrix}1 & r_{123} & r_{1j3}\\ r_{123} & 1 & r_{2j3}\\ r_{1j3} & r_{2j3} & 1\end{pmatrix}\right). \tag{A.6}$$

Annexe B

Codes R

B.1 Fonction de generation de donnees de type SNP

```
################################################
 Fonction qui simule m SNP pour n individus dont la frequence de
## l'allele mineure est donnee par f et la matrice V contient les
## correlation entre les SNP
#############
## Entree de la fonction SNP012
# n Nombre d'individus
# f la frequence de l'allele mineure du SNP (f<=0.5)
# V Matrice de correlation entre les SNP
## La fonction retourne un tableau dont les elements
## sont 0 ou 1 ou 2 qui est le nombre de fois que
##l'allele mineure est present dans le SNP.
SNP012 = function(n, f, V)
{
# Test si la frequence des alleles mineures sont bien
#inferieur a 0.5
 if((max(f)>0.5) || (min(f) < 0))
 {
 stop("The frequency of the minor allele must be
 positive and smaller than or equal to 0.5")
  }
  # Test si le nombre de SNP est en adequation avec la
  # matrice V donnee en entree
  m = length(f) # Nombre de SNP
  if( m != dim(V)[2])
```

```
  {
    stop("Dimensions are not adequate")
  }

  require(mnormt)
# On charge le ce package pour la loi normale multivariee.
  # Simulation de variable de loi normale multivariee
  #avec une moyenne nulle et une matrice de variance
  #covariance V fournie en entree
  Z = rmnorm(n, rep(0, m),  V)
  # On discretise ces variables par les quantiles de la
  #loi normale
  U = pnorm(Z) # Donc 0 <= U <= 1
  ## Construction des donnees de SNP = 0, 1, ou 2 pour
  #le nombre d'allele mineure decompte
  SNP = matrix(NA, n, m)
  # Soit A le nombre d'allele mineure.
  # La frequence de l'allele mineure f fournie en
  # entree
  # P(A = 0) = (1-f)^2
  # P(A = 1) = 2f(1-f)
  # P(A = 2) = f^2.
  for (i in 1:n) {
    for (j in 1:m) {
# Si U < P(A = 0)
      if(U[i,j] <= (1-f[j])^2 )
        # Alors SNP = 0
        SNP[i,j] = 0
 # Si  (1-f)^2 < U <= (1-f)^2 + 2f(1-f) = 1-f^2
      if((U[i,j] > (1-f[j])^2 ) & (U[i,j] < (1-f[j]^2)))
# Alors SNP = 1
        SNP[i,j] = 1
 # Si U >= (1-f)^2 + 2f(1-f) = 1-f^2
      if(U[i,j] >= (1-f[j]^2) )
 # Alors SNP = 2
        SNP[i,j] = 2
    }
  }
#Retourne les donnees sur les SNP sous forme d'un
```

```
#data.frame
 return(data.frame(SNP))
}
##########################################
```

B.2 Fonction d'approximation Monte Carlo de A_{1jk} et A_{2jk}

```
################################
# Approximation de Monte Carlo unidimensionnelle k=1
################################
## Entree de la fonction Integral_MC1
################################
## Z la statistique de test du SNP
## B le nombre de repetition pour MC
## sigmalambda l'erreur de la loi sur le NCP note
# lambda
## Cette fonction retourne les valeurs des integrales
## A11 et A12
Integral_MC_1 = function(Z, sigmalambda, N, B = 1000)
{
## Structre de controle des entrees de la fonction
##Integral_MC_1
 if(missing(Z))
  stop("Test statistics not provided.")
 ################################################
 # Loi normale sur la variable d'integration
 l1 = rnorm(B, 0, sigmalambda)
  D = NULL
  for (i in 1:B) {
    M1 = sqrt(N)*l1[i]
    V1 = 1+0.5*(l1[i])^2
     # Densite de la normale
    fZ = dnorm(Z, mean = M1, sd = sqrt(V1))
    D[i] = fZ
  }
  A1 = mean(D) # Estimateur MC de A11 (lambda != 0)
  A2 = dnorm(Z, mean = 0, sd = 1) # Valeur de A12 si lambda = 0

  return(list(A1 = A1, A2 = A2))
```

```
}
#########################
# Approximation de Monte Carlo pour k>1
################
## Entr e de la fonction Integral_MCk
#################################
## Z le vecteur des statistiques de test pour k SNP
## B le nombre de repetition pour MC
## sigmalambda l'erreur de la loi sur les NCP des k SNP
## Cette fonction retourne les valeurs des integrales
##Ak1 et Ak2
Integral_MC_k=function(Z,sigmalambda,Vk,N,B=1000)
{
## Structre de controle des entrees de la fonction
##Integral_MC_k
 if(missing(Z))
    stop("Test statistics not provided.")
  if(missing(Vk))
    stop("The V matrix not provided.")
  if(missing(N))
    stop("Sample size not provided.")
 # calcul de la dimension de l'integral qui correspond
 #a l'etape de selection
  d = dim(Vk)
  k = d[1]
  ##################################################
  MCV = matrix(NA, k, k)
  MCV0 = matrix(NA, k, k)
  #simule k variables (lambdas) de loi normale indep.
  m.MC = rep(0, k)
  mcv.MC = diag(sigmalambda^2, k,k)
  # NCP: parametre non central
  l = rmnorm(n = B, mean = m.MC, varcov = mcv.MC)
  l0 = l
  l0[1:B,k] = 0
  ##################
   D1 = NULL
   D2 = NULL
   Dk = (diag(Vk))^{-1}
```

```
   for (i in 1:B) {
    # Moyenne de Z
    M = sqrt(N*diag(Dk, k,k))%*%(l[i,])
    # Moyenne de Z si lambda.k = 0
    M0 = sqrt(N*diag(Dk, k,k))%*%(l0[i,])
   # Matrice de variance covariance de Z
MCV = sqrt(diag(Dk, k,k))%*%(Vk+0.5*((l[i,])
%*%t(l[i,])))%*%t(sqrt(diag(Dk, k,k)))
# Matrice de variance covariance si lambda.k = 0
MCV0 = sqrt(diag(Dk, k,k))%*%(Vk+0.5*((l0[i,])
%*%t(l0[i,])))%*%t(sqrt(diag(Dk, k,k)))
 # Densit  de Z
 # si lambdaj != 0
 fZ1 = dmnorm(t(Z), t(M), varcov = MCV)
 # si lambdaj = 0
 fZ2 = dmnorm(t(Z), t(M0), varcov = MCV0)
    D1[i] = fZ1
    D2[i] = fZ2
    }
 # Estimateur MC int gral Ak1
  A1 = mean(D1)
 # Estimateur MC int gral Ak2
  A2 = mean(D2)
  return(list(A1 = A1, A2 = A2))
}
##############################
```

B.3 Fonction de selection de SNPs

```
################################################
### Fonction qui selectionne k0 SNP causant
####### Indices utilisees
## i indice sur l'individu
## j indice sur le SNP
## k indice sur l'\'etape
###############################################
####### Entree de la fonction select.SNP ####
###############################################
##### y donnees sur le trait de longueur N
```

```
##### X donnees sur les SNP de dimension N*m
##### prob.p donnnees sur les proba a priori de
### longueur m
##### sigmalambda donnees sur l'erreur de la loi sur
## les NCP note lambda
##### k0 le nombre de SNP qu'on veux selectionner par
## defaut k0=1
##### B le nombre de repetition pour l'approximation
### des integrales par Monte Carlo
##### B = 1000 (pour avoir un estimateur convergent)
##########################################
####### Sortie de la fonction select.SNP #####
##########################################
##### indice_snp.c indice des SNP selectionnes
##### prob_post donne les proba a posteriori calcule
## des SNP a chaque etape
##### list_lambda.hat donne l'estimateur des NCP
### calcule des SNPs a chaque etape
##### Z.hat donne la statistique de test estimee des
## SNP a chaque etape
select.SNP=function(y,X,prob.p=NULL,sigmalambda=NULL,k0
=NULL,B=1000)
{
## Structre de controle des entrees de la fonction
##select.SNP
 if(missing(y))
  stop("Phenotype data not provided.")
  if(missing(X))
  stop("Data on SNPs not provided.")
  if(missing(k0))
  {
  warning("The number k0 of causal SNPs to be selected
  not provided(k0=1)")
    k0 = 1
  }

  ## Longueur de y, N et le nombre de SNPs m.
  d = dim(X)
  N = d[1] # Nombre d'individus
```

```
  m = d[2] # Nombre de SNP
  if((k0 <= 0) || (k0 > m ))
  stop("The number of SNPs can not be negative or
  greater than the number of existing SNPs")
## V\'erification si les information a priori sont
## fournies
  if(missing(prob.p))
  {
   #Proba a priori inconnues
 warning("Prior probabilities are not provided (prob.p = 0.5).")
 # tous les SNP ont une proba a priori egale 0.5
  prob.p = rep(0.5, m)
  }
  if(missing(sigmalambda))
  {
 #Ecart-type de la loi a priori des NCP inconnus
 warning("Standard deviations of the NCP prior law not
 provided (sigmalambda = 0.1).")
 # tous ecart-type des SNP initialises a 0.1
    sigmalambda = rep(0.1, m)
  }

  if(length(y) != N)
  stop("Dimensions are not adequate")

  if(length(prob.p) != length(sigmalambda))
    stop("Dimensions are not adequate")

  if(length(prob.p) !=m)
    stop("Dimensions are not adequate")

  if(length(sigmalambda) != m)
    stop("Dimensions are not adequate")

  if(all(is.na(y)))
  stop("Missing phenotype data")

  if( (prob.p < 0) || (sigmalambda < 0) )
stop("The probabilities or standard deviation can not
```

```
be negative")
 ######### Definitions des variables de sortie
  list_lambda.hat = list()
  Z.hat = list()
  prob_post = list()
  indice_snp.c = NULL
 ############## ETAPE
 ################################
  beta.hat1 = NULL
  Z1.hat = NULL
  prob.q1 = NULL
  list1_lambda.hat = rep(NA, m)
# Selection du premier SNP causant
  for (j in 1:m) {
########################################
# Regression individuelle sur les SNP X_j (j=1,...,m)
# Y = b_0 + b_j X_j + E
fit.X1 = lm(y ~ X[,j])
  beta.hat1 =  (summary(fit.X1)$coefficients)[,1]
sigma.hat1 = summary(fit.X1)$sigma
################################
## Estimation des NCP
  lambda.hat = beta.hat1[2]/sigma.hat1
## Statistique de Wald
 Z1.hat[j] = sqrt(N)*lambda.hat
 ###################################
 ## Calcul des integrales unidimensionnelles
A_Int=Integral_MC_1(N=N,Z=Z1.hat[j],B=B,sigmalambda=sigmalambda[j])
# Aprroximation MC de A11 (lambdaj != 0)
A11=A_Int$A1
# Aprroximation MC de A12 (lambdaj = 0)
A12 = A_Int$A2
###################################
# Proba a posteriori que lambda1 != 0
prob.q1[j] = ((1-prob.p[j])*A11)/(prob.p[j]*A12+(1-prob.p[j])*A11)
# Vecteur des estimes de NCP
    list1_lambda.hat[j] = lambda.hat
  }
# SNP causant == X_C1
```

```
 indice_snp.c[1] = which.max(prob.q1)
 # list prob a posteriori
 prob_post[[1]] = prob.q1
 # Estimation des NCP a l'etape 1
 list_lambda.hat[[1]] = list1_lambda.hat
 # Statistiques d'association a l'etape 1
  Z.hat[[1]] = Z1.hat
########################################
############ ETAPE suivantes k>1
if(k0>1)
{
  # Recherche des k0-1 autre SNP causant
  for (k in 2:k0) {
  listk_lambda.hat = matrix(NA, m, k0)
  list_Zk.hat = matrix(NA, m, k0)
  beta.hat = NULL
  Zk.hat = NULL
  prob.qk = NULL
  lambda.hat = NULL
 # Initialise a zero les proba a posteriori des SNP
 ## deja selectionnes
 # pour ne pas les selectionnes aux etapes suivantes
 prob.qk[indice_snp.c] = 0
 # Calcul le nombre de SNP selectionnes
 nb.c = length(indice_snp.c)
 # Calcul indice max pour beta.hat_k
 nb.c1 = nb.c+2
 for (j in 1:m){
 if(all(j != indice_snp.c)){
########################################
# Regression sur les autres SNP X_j non selectionnes
#(j != indice_snp.c)
      fit.X = lm(y~X[,indice_snp.c]+X[,j])
      beta.hat =  (summary(fit.X)$coefficients)[,1]
      sigma.hat = summary(fit.X)$sigma
 ########################################
 # Definition de la matrice schema Xk avec l'intercept
      Xk = cbind(X[,indice_snp.c], X[,j])
 # Calcul du nombre SNP sur Xk
```

```
      s = dim(Xk)[2]
# Calcul de la matrice V_k de l'etape k (defini dans le
##document)
    Vk = (N*solve(t(Xk)%*%Xk))
    Dk = (diag(Vk))^{-1}
    ## Estimation des NCP
    lambda.hat = beta.hat[2:nb.c1]/sigma.hat
    # Statistique de test Z
    Zk.hat = sqrt(N*diag(Dk, s,s))%*%lambda.hat
   # A priori sur la loi de lambda
sigmalambdaj=c(sigmalambda[indice_snp.c],sigmalambda[j])
########################################
 # Calcul des integrales par M.C.
 A_Int=Integral_MC_k(N=N,Z=Zk.hat[,1],Vk
 =Vk,B=B,sigmalambda=sigmalambdaj)
 # Approximation MC de Ak1 (lambdaj != 0)
   Ak1 = A_Int$A1
 # Approximation MC de Ak2 (lambdaj = 0)
   Ak2 = A_Int$A2
# Proba a posteriori que lambdak != 0
prob.qk[j] = (prod((1-prob.p[indice_snp.c]))*(1-prob.p[j])*Ak1)/
   (prod((1-prob.p[indice_snp.c]))*(1-prob.p[j])*Ak1+
    prod((1-prob.p[indice_snp.c]))*prob.p[j]*Ak2)
  # Estimation des NCP a l'etape k
listk_lambda.hat[j,1:length(lambda.hat)] = lambda.hat
# Statistiques de test a l'etape k
 list_Zk.hat[j,1:length(lambda.hat)] = Zk.hat
    }
  }
 # SNP causant == X_Ck
 indice_snp.c[k] = which.max(prob.qk)
# list prob a posteriori
prob_post[[k]] = prob.qk
# Initialisation des estimateurs des NCP a l'etape k
 list_lambda.hat[[k]] = listk_lambda.hat
# Initialisation des statistiques de test a l'etape k
 Z.hat[[k]] = list_Zk.hat
 }
}
```

```
return(list(indice_snp.c = indice_snp.c, prob_post = prob_post,
list_lambda.hat = list_lambda.hat, Z.hat = Z.hat))
}
########################################
```

Bibliographie

[1] Genetics home reference. what are single nucleotide polymorphisms (snps) ?, 2020. URL https://ghr.nlm.nih.gov/primer/genomicresearch/snp.

[2] C. Beroud. *Bases de données et outils bioinformatiques utiles en génétique*. Collège National des Enseignants et Praticiens de Génétique Médicale, 2010–2011.

[3] Adolf Buse. The likelihood ratio, wald, and lagrange multiplier tests : An expository note. *The Amercican Statistician*, 36(3) :Part 1 153–157, August 1982. doi : 10.1080/00031305.1982.10482817.

[4] Charles C. Chung. Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer. *Hum. Mol. Genet*, 20(14) : 2869–2878, July 1982. doi : 10.1093/hmg/ddr189.

[5] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491 :56–65, November 2012.

[6] Ashley H. Beecham et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11) :1353–1360, 2013. doi : 10.1038/ng.2770.

[7] Christian Benner et al. Finemap : Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10) :1493–1501, May 2016. doi : 10.1093/bioinformatics/btw018.

[8] David E. Reich et al. Linkage disequilibrium in the human genome. *Nature*, 441(6834) : 199–204, May 2001. doi : 10.1093/bioinformatics/btw018.

[9] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414) :57–74, 2012.

[10] Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11) :1274–1283, 2013. doi : 110.1038/ng.2797.

[11] Hakon Hakonarson et al. A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. *Nature*, 448(7153) :591–594, July 2007.

[12] Hana Lango Allen et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317) :832–838, 2010. doi : 10.1038/nature09410.

[13] Julian B. Maller et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12) :1294–1301, Oct 2012. doi : https://doi.org/10.1038/ng.2435.

[14] Michael J. Flister et al. Identifying multiple causative genes at a single gwas locus. *Genome Res.*, 23(12) :1996–2002, Dec 2013. doi : 10.1101/gr.160283.113.

[15] Miriam S. Udler et al. Fgfr2 variants and breast cancer risk : Fine-scale mapping using african american studies and analysis of chromatin conformation. *Hum. Mol. Genet.*, 18 (9) :1692–1703, Feb 2009. doi : 10.1093/hmg/ddp078.

[16] Rhett M. Schiffman et al. Utility assessment among patients with dry eye disease. *Ophthalmology*, 110(7) :1412–1419, July 2003. doi : https://doi.org/10.1016/S0161-6420(03)00462-7.

[17] Stephan Ripke et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, 45(10) :1150–1159, August 2013.

[18] Wenan Chen et al. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3) :719–736, July 2015. doi : 10.1534/genetics.115.176107.

[19] Farhad Hormozdiari E Kostem EY Kang, Bogdan Pasaniuc and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2) : 497–508, Oct 2014. doi : 10.1534/genetics.114.167908.

[20] Mickaël Guedj. *Méthodes Statistiques pour l'Analyse de Données Génétiques d'Association à Grande Echelle*. PhD thesis, Université d'Evry-Val d'Essonne, 2007.

[21] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in transethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2) :260–271, August 2015. doi : 10.1016/j.ajhg.2015.06.007.

[22] Aurélie Labbe. De la statistique à la génétique : identifier les gènes responsables de maladies complexes. *Bulletin AMQ*, LIII(2), May 2013.

[23] David Altshuler Mark J. Daly and Eric S. Lander. Genetic mapping in human disease. *Science*, 322(5903) :881–888, November 2008. doi : 10.1126/science.1156409.

[24] Etienne Muller. *Les défis du séquen age à haut débit dans l'exploration génétique des cancers du sein et de l'ovaire*. PhD thesis, Université de Rouen, 2018.

[25] Jonathan K. Pritchard and Molly Przeworski. Linkage disequilibrium in humans : Models and data. *The American Journal of Human Genetics*, 68(1) :1–14, Jun 2001. doi : 10.1086/321275.

[26] Rainard B. Robbins. Some applications of mathematics to breeding problems iii. *Genetics*, 3(4) :375–389, July 1918.

[27] Sara Lindstr m Peter Kraft Gleb Kichaev Mergan Roytman Ruth Johnson, Eleazar Eskin and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2) :248–255, 2017. doi : 10.1093/bioinformatics/btw615.

[28] Eleazar Eskin Alkes L Price Peter Kraft Gleb Kichaev Wen-Yu Yang Sara Lindstr m, Farhad Hormozdiari and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10), Oct 2014. doi : 10.1371/journal.pgen.1004722.

[29] Elizabeth A. Thompson. *Statistical Inference from Genetic Data on Pedigrees*, volume 6. Institute of Mathematical Statistics, 2000.

[30] Farhad Hormozdiari Gleb Kichaev Wen-Yu Yang, Bogdan Pasaniuc and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 36(12) :i206–i213, 2015. doi : 10.1093/bioinformatics/btv240.

[31] Andreas Ziegler and Inke R. K nig. *A Statistical Approach to Genetic Epidemiology*. WILEY, 2010. ISBN 10.1002/9783527633654.