UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TEST D'ASSOCIATION BIVARIÉ POUR PHÉNOTYPES NON NORMAUX

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

JULIEN ST-PIERRE

MARS 2018

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Les avancées récentes dans le séquençage nouvelle-génération de l'ADN ont permis d'identifier des millions de variants génétiques rares, jusqu'ici ignorés dans les études d'association génétique. Le but de ce projet est de proposer un nouveau modèle, basé sur les copules, afin de tester l'association entre plusieurs variants rares et/ou communs dans une région génomique et un phénotype bivarié continu. Puisque les copules permettent de modéliser la dépendance entre les traits indépendamment de leurs lois de répartitions marginales, on est en mesure de relaxer l'hypothèse de normalité marginale pour le phénotype bivarié. Afin d'intégrer les variants rares dans l'analyse d'association génétique, on propose un modèle de régression mixte pour lequel l'effet de la région génétique est supposé aléatoire. Ainsi, on teste l'association sur l'ensemble de la région à l'aide d'un test de score sur une composante de variance. Pour obtenir la statistique du test du score, il faut toutefois obtenir une forme explicite pour la vraisemblance du modèle. On propose de résoudre numériquement cette intégrale à l'aide de l'approximation de Laplace de la fonction de vraisemblance conditionnelle sous l'hypothèse nulle. Nous sommes en mesure de démontrer que la distribution théorique de notre statistique de score suit un mélange de lois de Chi-deux sous l'hypothèse nulle. À l'aide de simulations, nous démontrons que l'erreur de type I pour notre test est bien contrôlée pour divers scénarios. Les résultats obtenus démontrent aussi que la puissance de notre test est comparable aux tests d'association génétique MURAT et SKAT. Nous appliquons notre méthode sur les données génétiques réelles de l'étude ALSPAC.

Mots clés : copules, variants rares, test d'association multivarié, modèle linéaire généralisé mixte.

INTRODUCTION

La statistique génétique est une branche de la statistique qui s'intéresse plus spécifiquement à l'analyse des traits hérités et aux données génétiques. Le but de cette branche est de développer des méthodes statistiques qui permettent d'identifier les facteurs de risque génétiques pour des traits mesurables ou des maladies complexes. Les études de type GWAS (Genome-wide association study) ont permis d'identifier des centaines de variants génétiques communs associés à des traits complexes. Toutefois, ces études ont mis en lumière le problème du manque d'héritabilité, c'est-à-dire que les variants génétiques identifiés par GWAS n'expliquent qu'une faible partie de la variabilité totale observée pour les traits étudiés. Pour remédier à ce problème, il faut considérer d'autres modèles que celui supposant que les traits communs sont influencés seulement par des variants communs dans la population.

Les avancées récentes dans le séquençage nouvelle-génération de l'ADN ont permis d'identifier des millions de variants génétiques rares, jusqu'ici ignorés dans les études d'association génétique. La découverte de ces variants rares a nécessité le dévelopemment de nouvelles méthodes statistiques, dont notamment le test SKAT (Wu *et al.*, 2011), qui permet de tester l'association entre un trait quantitatif ou binaire et une région génétique composée de plusieurs variants rares et/ou communs. Dans le but d'augmenter la puissance statistique de ces tests, on peut tester l'association entre une région génétique et plusieurs traits simultanément, c'est-à-dire une variable dépendante multivariée. Toutefois, la plupart des tests multivariés, dont le test de MURAT (Sun *et al.*, 2016), présument une dépendance linéaire entre les différents traits ainsi qu'une distribution multivariée normale

pour la variable dépendante. En pratique, on travaille souvent avec des traits qui ne sont pas normalement distribués, d'où le besoin de développer un nouveau test d'association entre variants rares et phénotypes multivariés de loi non normale. Dans ce projet, on propose un nouveau test multivarié basé sur les copules , CBM-RV (Copula Based Multivariate Rare Variants), pour identifier l'association entre plusieurs variants rares dans une région génétique et un phénotype bivarié continu.

Dans le premier chapitre, on présente quelques concepts de base nécessaires pour la compréhension du développement théorique derrière notre méthode. Tout d'abord, on passe en revue quelques concepts de base sur la génétique et l'hérédité. Ensuite, on introduit les différents modèles de régression couramment utilisés en statistique avant de faire quelques rappels sur l'estimation et les tests d'hypothèses basés sur la vraisemblance. Dans le deuxième chapitre, on présente la théorie des copules ainsi que les différentes familles qui sont utilisées pour modéliser la dépendance entre deux ou plusieurs traits. Dans le troisième chapitre, on présente le développement théorique pour le test CBM-RV que nous proposons. Dans le quatrième chapitre, des études de simulations permettent de démontrer la performance de notre méthode. Enfin, dans le cinquième chapitre on applique le test CBM-RV sur des données réelles provenant de la cohorte ALSPAC .

CHAPITRE I

CONCEPTS DE BASE

1.1 Concepts de base en génétique

Avant d'introduire des notions mathématiques, on présente dans cette section quelques concepts de base en génétique qui permettent de mieux comprendre la nature des données que nous modélisons avec notre test d'association.

1.1.1 ADN, gène, chromosome

Le matériel génétique de chaque individu est retrouvé dans une molécule appelée acide désoxyribonucléique (ADN), qui est présente dans toutes les cellules du corps humain et est la plupart du temps retrouvée sous la forme d'une double hélice, tel que présenté dans la Fig. 1.1. Les molécules d'ADN sont un arrangement de plus petites molécules, ou monomères, appelées nucléotides. Chaque nucléotide contient un sucre appelé désoxyribose, un groupement phosphate et une base azotée. Étonnemment, il n'existe que 4 bases azotées différentes dans les molécules d'ADN, soient l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C). De plus, les deux brins formant la double hélice d'ADN sont liés entre eux par un pont hydrogène entre leurs bases azotées respectives suivant une règle d'appariement stricte : A avec T et C avec G. Les deux brins d'ADN sont donc complémentaires.

Chaque cellule du corps humain contient, dans son noyau, les chromosomes qui sont le support physique du matériel génétique. À l'exception des cellules sexuelles, toutes les cellules contiennent 23 paires de chromosomes. C'est sur les chromosomes que l'on retrouve les gènes, qui sont l'unité de base de l'hérédité. Les gènes ne sont rien de plus qu'une séquence d'ADN particulière, délimitée au début et à la fin par des séquences particulières de nucléotides, appelées exons. Les gènes contiennent les instructions qui permettent de fabriquer des protéines spécifiques. En effet, chaque triplet de nucléotides sur un gène correspond à un acide aminé particulier, qui est en quelque sorte l'unité constituante des protéines. Les protéines jouent un rôle crucial dans tout le corps puisqu'elles interviennent dans presque toutes les fonctions cellulaires. Ces fonctions incluent notamment l'accélération des réactions chimiques, le transport de substances, la mise en réserve d'acides aminés et la protection contre les maladies. L'ensemble des gènes d'un individu, le génotype, affecte les caractères physiques et physiologiques, le phénotype, par le biais des protéines.

1.1.2 Diversité et expression des gènes

La variation observée pour certains phénotypes dans une population est attribuable à l'environnement et aux variations des gènes qui codent pour ces phénotype, à différents degrés d'importance selon le caractère phénotypique étudié. Par exemple, l'ADN situé sur le locus du gène qui code pour la couleur des yeux est différent d'un sujet à l'autre. Dépendamment de la séquence précise de nucléotides à ce locus, un individu peut avoir les yeux bleus ou verts. Ces variations de nucléotides pour un locus précis sont appelées allèles. Étant donné que chaque individu hérite de paires de chromosomes homologues à la naissance, un chromosome provenant du père et un provenant de la mère, il y précisément deux allèles pour chaque caractère. Si les deux allèles d'un locus sont identiques pour un caractère donné, on parle d'un individu qui est homozygote pour ce caractère. Un sujet qui

possède deux allèles différents pour ce caractère est dit hétérozygote. Puisqu'un individu reçoit aléatoirement un allèle par parent indépendamment d'un caractère à l'autre, on observe une variation génétique importante chez les descendants d'une même famille.

La plus grande partie de la diversité génétique observée est due aux variations génétiques ne touchant qu'une seule paire de bases nucléotidiques, appelées polymorphismes nucléotidiques simples, ou SNP en anglais. L'allèle le plus présent dans une population pour un SNP est appelé allèle majeur par opposition à l'allèle mineur, qui est le moins fréquent. On définit un variant génétique rare comme un SNP pour lequel la fréquence allélique mineure (MAF) est inférieure à 5%. Dû à leur faible représentation dans la population, les variants rares auraient un impact fonctionnel beaucoup plus important que les variants communs, d'où l'intérêt de les inclure dans les études d'association.

Il existe différents modes d'expression des gènes, dont la dominance, la pléitropie, la polymérie et l'épistasie. Les phénomènes de dominance font référence aux conséquences de l'interaction entre différents allèles homologues pour un caractère donné. Par exemple, chez la souris, les individus homozygotes CC et hétérozygotes Cc pour le locus déterminant le caractère de la coloration du pelage seront colorés, tandis que les individus homozygotes cc seront albinos. Les individus hétérozygotes et homozygotes CC sont donc indiscernables pour ce caractère, étant donné que l'allèle C est dominant et que l'allèle c est récessif. Celui-ci est exprimé seulement dans le cas où l'allèle dominant est absent. On parle de codominance dans le cas où, pour les individus hétérozygotes, les deux allèles déterminent simultanément le caractère exprimé.

La pléitropie fait référence aux situations pour lesquelles une mutation dans un gène détermine plusieurs caractères qui semblent à priori indépendants. Par

Figure 1.1: Molécule d'ADN

exemple, il existe chez le hamster un gène qui détermine à la fois l'aspect blanc du pelage et l'aspect réduit des yeux. La polymérie est en quelque sorte la situation inverse, c'est-à-dire qu'on observe plusieurs gènes différents avoir un impact sur le même caractère. Enfin, on parle d'épistasie lorsque des phénomènes de dominance existent entre gènes codant pour différents caractères. Un gène A qui influence ou masque l'expression d'un autre gène B, codant pour un phénotype différent, est dit épistatique sur le gène B.

1.2 Notions d'algèbre linéaire

Le test d'association génétique que nous proposons est applicable pour des variables dépendantes multivariées, c'est-à-dire de dimension supérieure à un. Il est donc important de rappeler dans cette section quelques notions d'algèbre linéaire qui sont nécessaires pour la compréhension de l'analyse statistique multivariée.

a. Soient $\boldsymbol{A}$ et $\boldsymbol{B}$ deux matrices carrées d'ordre n, $\boldsymbol{C}$ une matrice de taille $n \times p$,

$\boldsymbol{D}$ une matrice de taille $p \times n$ et soit la trace de $\boldsymbol{A}$ définie par $\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{ii}$, alors on a

$$\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B}), \tag{1.1}$$

$$\text{tr}(\boldsymbol{CD}) = \text{tr}(\boldsymbol{DC}). \tag{1.2}$$

b. Soit la forme quadratique $\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}$, avec $\boldsymbol{\beta}$ un vecteur aléatoire de $\mathbb{R}^n$ et $\boldsymbol{A}$ une matrice de coefficients carrée d'ordre n, l'espérance est donnée par

$$\begin{aligned} \text{E}[\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}] &= \text{E}[\text{tr}(\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta})] \\ &= \text{E}[\text{tr}(\boldsymbol{A} \boldsymbol{\beta} \boldsymbol{\beta}^T)] \text{ par (1.2)} \\ &= \text{tr}(\text{E}[\boldsymbol{A} \boldsymbol{\beta} \boldsymbol{\beta}^T]) \\ &= \text{tr}(\boldsymbol{A}\ \text{E}[\boldsymbol{\beta} \boldsymbol{\beta}^T]). \end{aligned} \tag{1.3}$$

c. Un vecteur non nul $\boldsymbol{u}$ est un vecteur propre de la matrice carrée $\boldsymbol{A}$ s'il satisfait l'équation

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}\ ,$$

où λ est un scalaire appelé valeur propre de $\boldsymbol{A}$. De plus, si $\boldsymbol{A}$ est une matrice non singulière de taille $n \times n$, alors il existe n valeurs propres et n vecteurs propres non nuls de $\boldsymbol{A}$.

d. Pour toute matrice carrée $\boldsymbol{A}$ qui possède n valeurs propres λ_1, λ_2, ... , λ_n, on a

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i, \tag{1.4}$$

$$|A| = \prod_{i=1}^{n} \lambda_i,$$

où |A| est défini comme le déterminant de la matrice A.

e. Soit $\boldsymbol{A}$ une matrice symétrique d'ordre n, alors ses vecteurs propres sont mutuellement orthogonaux, i.e.

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = 0 \text{ pour } i = 1, ..., n \text{ et } i \neq j.$$

De plus, soit $\boldsymbol{U}$ la matrice formée par les vecteurs propres normalisés de $\boldsymbol{A}$, alors $\boldsymbol{U}$ est orthogonale, i.e.

$$\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_n,$$

avec $\boldsymbol{I}_n$ la matrice identité d'ordre n.

f. Soient $\boldsymbol{A}$ une matrice symétrique d'ordre n et $\boldsymbol{D}$ une matrice diagonale d'ordre n contenant les valeurs propres de $\boldsymbol{A}$, alors la décomposition spectrale de $\boldsymbol{A}$ est donnée par

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T. \tag{1.5}$$

g. Soient $\boldsymbol{A}$ une matrice de taille $n \times m$ et $\boldsymbol{B}$ une matrice de taille $p \times q$, alors le produit de Kronecker $\boldsymbol{A} \otimes \boldsymbol{B}$ est défini par

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} A_{11}\boldsymbol{B} & \dots & A_{1m}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ A_{n1}\boldsymbol{B} & \dots & A_{nm}\boldsymbol{B} \end{bmatrix}$$

où $\boldsymbol{A} \otimes \boldsymbol{B}$ est une matrice de taille $np \times mq$.

De plus, on a

$$\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}),$$

$$(\boldsymbol{A} \otimes \boldsymbol{B})^T = \boldsymbol{A}^T \otimes \boldsymbol{B}^T.$$

1.3 Modèles de régression

En épidémiologie, on cherche souvent à modéliser la relation entre une variable d'intérêt et une ou plusieurs covariables. Par exemple, on pourrait vouloir étudier l'effet de la consommation d'alcool sur la pression sanguine, ou encore l'effet de certaines variations génétiques sur l'incidence d'une maladie. Notre test d'association génétique étant basé sur un modèle de régression linéaire généralisé mixte, il convient de rappeler la théorie des différents modèles de régression existants.

Dans cette section, on présente donc deux types de modèles importants, soient les modèles linéaires généralisés et les modèles linéaires généralisés mixtes.

1.3.1 Modèles linéaires généralisés

Si on travaille avec N observations indépendantes d'une même variable réponse Y, distribuée selon une loi normale, alors le modèle linéaire à effets fixes suppose que

$$\mathrm{E}[Y_i] = \mu_i = \boldsymbol{X}_i^T \boldsymbol{\gamma} \tag{1.6}$$

et

$$\mathrm{Var}[Y_i] = \sigma^2,$$

où $\boldsymbol{\gamma}$ est le vecteur de coefficients représentant les effets fixes inconnus pour chacune des k covariables et pour l'ordonnée à l'origine s'il y a lieu, $\boldsymbol{X_i}$ est un vecteur contenant les observations des k covariables pour le $i^{\text{ème}}$ sujet, σ^2 est la variance résiduelle et $i = 1, ..., N$. Suivant ce modèle, on a que

$$Y_i \sim N(\boldsymbol{X_i^T \gamma}\ , \sigma^2). \tag{1.7}$$

Dans le cadre plus général des modèles linéaires généralisés, on va supposer que les observations y_i proviennent d'une loi de famille exponentielle, c'est-à-dire une distribution pour laquelle il est toujours possible d'écrire la fonction de densité telle que

$$f(y_i) = \exp[\frac{y_i\theta_i - b(\theta_i)}{\phi} - c(y_i, \phi)], \tag{1.8}$$

où θ_i est le paramètre naturel de la loi et ϕ le paramètre de dispersion. Dépendamment de la distribution suivie par Y, il peut arriver que la relation entre l'espérance de Y_i, μ_i, et le prédicteur $\boldsymbol{X_i^T \gamma}$ ne soit pas directement linéaire comme dans l'équation (1.6). Dans ce cas, on procède à une transformation de la moyenne μ_i en utilisant une fonction de lien $g(\cdot)$, de façon à obtenir la relation suivante :

$$g(\mu_i) = \boldsymbol{X_i^T} \gamma.$$

Une propriété importante des modèles linéaires généralisés est le lien particulier entre la fonction $b(\theta)$ et les deux premiers moments de la variable Y. Pour démontrer cette propriété, on note que, dans le cas où $f(y)$ est continue, nous avons que

$$\int_{-\infty}^{\infty} \exp[\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)]dy = 1$$

par définition de la fonction de densité. On dérive des deux côtés par rapport à θ, d'où

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} \exp[\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)]dy = \frac{\partial}{\partial\theta}1 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\exp[\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)]dy = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \exp[\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)] \cdot \frac{y - b'(\theta)}{\phi}dy = 0.
\end{aligned}$$

Par l'équation (1.8),

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{y - b'(\theta)}{\phi} f(y)dy = 0. \\
\Leftrightarrow &\int_{-\infty}^{\infty} y\ f(y)dy = b'(\theta)\int_{-\infty}^{\infty} f(y)dy \\
&\Leftrightarrow \mathrm{E}[Y] = b'(\theta). \qquad (1.9)
\end{aligned}$$

On obtient le deuxième moment de Y en calculant la dérivée seconde de $b(\theta)$, soit

$$\begin{aligned}
b''(\theta) &= \frac{\partial}{\partial\theta}\mathrm{E}[Y] \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} y\ f(y)dy \\
&= \int_{-\infty}^{\infty} y\ f(y) \cdot \frac{y - b'(\theta)}{\phi}dy.
\end{aligned}$$

On multiplie ϕ des deux côtés de l'équation et on développe l'intégrale,

$$\phi \cdot b''(\theta) = \int_{-\infty}^{\infty} y^2\ f(y)dy - b'(\theta)\int_{-\infty}^{\infty} y\ f(y)dy.$$

En utilisant le résultat (1.9),

$$\begin{aligned}\phi \cdot b''(\theta) &= E[Y^2] - (E[Y])^2 \\ &= \text{Var}[Y].\end{aligned}$$

En notant $v(\mu) = b''(\theta)$, on trouve

$$\phi \cdot v(\mu) = \text{Var}[Y], \tag{1.10}$$

avec $v(\mu)$ définie comme la fonction variance puisque qu'elle contient la partie de la variance qui dépend de la moyenne.

1.3.2 Modèles linéaires généralisés mixtes

Pour certains modèles, il n'est pas réaliste de supposer que l'effet d'une covariable sur la variable réponse est une constante. En effet, on peut supposer par exemple que l'effet de la consommation d'alcool sur la pression sanguine est très variable d'une personne à l'autre, c'est-à-dire qu'il y a une variabilité inter-sujet importante. Dans un modèle linéaire généralisé mixte, on va donc supposer que certaines covariables ont un effet aléatoire sur la variable réponse, traduisant la variabilité liée à chaque sujet, tandis que d'autres covariables ont un effet fixe, représentant l'effet de la population. Soit $\boldsymbol{\beta}$ le vecteur de coefficients de dimension r représentant les effets aléatoires inconnus de r covariables sur Y, alors on a

$$g(\text{E}[Y_i|\boldsymbol{\beta}]) = \boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta}, \tag{1.11}$$

où $\boldsymbol{Z_i}$ est le vecteur contenant les observations des covariables à effets aléatoires. Sans perte de généralité, on suppose que

$$\boldsymbol{\beta} \sim F(\mathbf{0}_r, \boldsymbol{\Sigma}_\beta).$$

Pour obtenir l'espérance marginale de Y_i, on utilise la propriété suivante :

$$\begin{aligned}\text{E}[Y_i] &= \text{E}[\text{E}[Y_i|\beta]] \\ &= \text{E}[g^{-1}(\boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta})], \text{ par (1.11).}\end{aligned} \tag{1.12}$$

Pour obtenir la variance marginale de Y_i, on utilise la propriété suivante :

$$\begin{aligned}\text{Var}[Y_i] &= \text{E}[\text{Var}[Y_i|\boldsymbol{\beta}]] + \text{Var}[\text{E}[Y_i|\boldsymbol{\beta}]] \\ &= \text{E}[\phi \cdot v(g^{-1}(\boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta}))] + \text{Var}[g^{-1}(\boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta})] \qquad (1.13)\end{aligned}$$

en utilisant les équations (1.10) et (1.11). De façon générale, il est difficile d'obtenir une forme simplifiée pour les équations (1.12) et (1.13) étant donné que g^{-1} est non linéaire, sauf dans le cas des modèles linéaires mixtes où la fonction de lien est la fonction identité. Dans ce cas, l'espérance marginale pour Y_i devient

$$\text{E}[Y_i] = \text{E}[\boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta}] = \text{E}[\boldsymbol{X}_i^T\boldsymbol{\gamma}]$$

puisque $\boldsymbol{\beta}$ est centré sur $\mathbf{0}$. La variance marginale pour Y_i est obtenue à l'aide de la relation suivante :

$$\begin{aligned}\text{Var}[Y_i] &= \text{E}[\text{Var}[Y_i|\boldsymbol{\beta}] + \text{Var}[\text{E}[Y_i|\boldsymbol{\beta}]] \\ &= \text{E}[\sigma^2] + \text{Var}[\boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{Z}_i^T\boldsymbol{\beta}], \text{ par (1.6) et (1.11)} \\ &= \sigma^2 + \text{Var}[\boldsymbol{Z}_i^T\boldsymbol{\beta}] \\ &= \sigma^2 + \boldsymbol{Z}_i^T\Sigma_{\boldsymbol{\beta}}\boldsymbol{Z}_i.\end{aligned}$$

Suivant le modèle linéaire mixte, on a

$$Y_i \sim N(\boldsymbol{X}_i^T\boldsymbol{\gamma}, \sigma^2 + \boldsymbol{Z}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{Z}_i),$$

c'est-à-dire que les effets fixes affectent seulement la moyenne tandis que les effets aléatoires affectent seulement la variance de Y_i.

1.4 Vraisemblance

Étant donné que la méthodologie que nous présentons dans ce mémoire se base sur un test du score, il convient de faire quelques rappels théoriques concernant l'inférence statistique basée sur la fonction de vraisemblance. On présente un

estimateur basé sur la fonction de vraisemblance avant d'introduire un test d'hypothèse couramment rencontré en statistique, soit le test du score de Rao (Rao, 2001).

1.4.1 Estimateur du maximum de vraisemblance

Soient $Y_1, Y_2, ..., Y_n$ des variables aléatoires dont la fonction de densité conjointe est donnée par $f(y_1, y_2, ..., y_n|\boldsymbol{\theta})$. Alors, la fonction de vraisemblance pour le vecteur de paramètres $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ est

$$L(\boldsymbol{\theta}) = f(y_1, ..., y_n|\boldsymbol{\theta}).$$

L'estimateur du maximum de vraisemblance pour $\boldsymbol{\theta}$, noté $\hat{\boldsymbol{\theta}}_{MLE}$, est donné par la valeur de $\boldsymbol{\theta}$ qui maximise la fonction de vraisemblance du modèle, c'est-à-dire qui maximise la probabilité des données observées. En pratique, il peut être plus facile de maximiser le logarithme naturel de la fonction de vraisemblance, ce qui est équivalent étant donné qu'il s'agit d'une fontion strictement monotone. Ainsi, la log-vraisemblance pour $\boldsymbol{\theta}$ est notée

$$l(\boldsymbol{\theta}) = \log f(y_1, ..., y_n|\boldsymbol{\theta}).$$

Dans le cas où les observations y_i sont indépendantes et identiquement distribuées (i.i.d.), on obtient que

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}).$$

On peut montrer, sous certaines conditions, que $\hat{\boldsymbol{\theta}}_{MLE}$ est un estimateur convergent de la vraie valeur $\boldsymbol{\theta}_0$, c'est-à-dire que lorsque n tend vers l'infini, $\hat{\boldsymbol{\theta}}_{MLE}$ converge vers $\boldsymbol{\theta}_0$. De plus, la distribution asymptotique de l'estimateur du maximum de vraisemblance est approximativement normale, tel qu'énoncé dans le Théorème 1.4.1.

THÉORÈME 1.4.1 — *Lorsque* $n \to \infty$,

$$\boldsymbol{I}(\boldsymbol{\theta}_0)^{1/2}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0) \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p),$$

avec $\boldsymbol{I}(\boldsymbol{\theta})$ *la matrice d'information de Fisher définie telle que*

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\,[\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^T}]. \tag{1.14}$$

Enfin, on présente dans l'annexe A un rappel théorique concernant la dérivabilité d'une fonction par rapport à un vecteur.

1.4.2 Test du score

Dans le but de tester l'hypothèse nulle simple $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$, il est possible de faire appel au test du score de Rao. Un avantage d'utiliser le test du score pour tester $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ est que ce test ne requiert aucune estimation de $\boldsymbol{\theta}$, ce qui peut s'avérer complexe dans un modèle avec effets aléatoires. De plus, le test du score est le test le plus puissant lorsque la vraie valeur de $\boldsymbol{\theta}$ est proche de $\boldsymbol{\theta}_0$. Soit le vecteur de paramètres $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$, la statistique pour le test du score sous $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ est donné par

$$\frac{\partial l}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \underset{H_0}{\sim} N_p(\mathbf{0}_p, I(\boldsymbol{\theta})). \tag{1.15}$$

Dans plusieurs cas, on cherche seulement à tester un sous-ensemble des paramètres du modèle, soient les paramètres d'intérêt. On définit alors les paramètres de nuisance du modèle comme des paramètres à estimer, mais dont on ne cherche pas à tester les valeurs. Toutefois, il faut tenir compte de l'estimation des paramètres de nuisance lorsqu'on applique le test du score sur les paramètres d'intérêt du modèle.

Par exemple, soit $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ la log-vraisemblance du modèle, qui dépend d'un vecteur de paramètres de nuisance $\boldsymbol{\theta} \in \mathbb{R}^m$ et d'un vecteur de paramètres d'intérêt $\boldsymbol{\beta} \in \mathbb{R}^k$, alors le vecteur score est donné par

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{bmatrix}.$$

On peut décomposer la matrice d'information observée de Fisher tel que

$$\begin{aligned} \boldsymbol{I}_{\text{obs}}(\boldsymbol{\beta},\boldsymbol{\theta}) &= -\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} \\ &= \begin{bmatrix} \boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) & \boldsymbol{I}_{\beta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) \\ \boldsymbol{I}_{\theta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) & \boldsymbol{I}_{\theta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) \end{bmatrix}, \end{aligned}$$

avec $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ et en notant son inverse par

$$\boldsymbol{I}_{\text{obs}}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1} = \begin{bmatrix} \boldsymbol{I}^{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) & \boldsymbol{I}^{\beta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) \\ \boldsymbol{I}^{\theta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) & \boldsymbol{I}^{\theta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) \end{bmatrix}.$$

On peut montrer que (Lu et Shiou, 2000)

$$\boldsymbol{I}^{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) = \boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1} + \boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1}\boldsymbol{I}_{\beta\theta}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{Z}^{-1}\boldsymbol{I}_{\theta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1},$$

avec

$$\boldsymbol{Z} = \boldsymbol{I}_{\theta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) - \boldsymbol{I}_{\theta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1}\boldsymbol{I}_{\beta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}),$$

$$\boldsymbol{I}_{\beta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

$$\boldsymbol{I}_{\theta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T},$$

et

$$\boldsymbol{I}_{\beta\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) = \boldsymbol{I}_{\theta\beta}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}^T}.$$

La statistique pour le test du score sous $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ est donné par

$$\frac{\partial l}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}})^T\boldsymbol{I}^{\beta\beta}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}})\frac{\partial l}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}}) \underset{H_0}{\sim} \chi^2_k,$$

avec $\hat{\boldsymbol{\theta}}$ estimé sous H_0. En d'autres mots, on a

$$\frac{\partial l}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}}) \underset{H_0}{\sim} N_k(\mathbf{0}_k, [\boldsymbol{I}^{\beta\beta}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}})]^{-1}). \tag{1.16}$$

1.5 Test du score SKAT

On présente dans cette section un exemple d'application d'un test du score pour un modèle linéaire mixte dans le cadre d'une analyse d'association génétique. Supposons qu'on observe un trait Y_i pour N individus indépendants, $i = 1, ..., N$, m covariables, $\boldsymbol{X_i} = (x_{i1}, x_{i2}, ..., x_{im})^T$, et un ensemble de SNPs, $\boldsymbol{G_i} = (g_{i1}, g_{i2}, ..., g_{ir})^T$ contenant r variants bialléliques codés $0, 1, 2$, représentant le nombre d'allèles mineurs. On s'intéresse à l'association entre Y_i et $\boldsymbol{G_i}$ en supposant le modèle linéaire mixte (Wu *et al.*, 2011) tel que

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\gamma} + \boldsymbol{G}_i^T\boldsymbol{\beta} + \epsilon_i \text{ pour } i = 1, ..., N, \tag{1.17}$$

où $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_m)^T$ et $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)^T$ sont les coefficients de régression pour $\boldsymbol{X_i}$ et $\boldsymbol{G_i}$ respectivement. De plus, on suppose que $\boldsymbol{\gamma}$ est un vecteur d'effets fixes tandis que $\boldsymbol{\beta} \sim N(\boldsymbol{0}_r, \tau\boldsymbol{W})$ est un vecteur d'effets aléatoires, avec τ une composante de variance inconnue et $\boldsymbol{W} = \text{diag}(w_1, ..., w_r)$ une matrice de poids telle que w_j est le poids associé au j^e variant. Les poids w_j sont fixés à priori et sont, de manière générale, inversement proportionels à la fréquence allélique. Finalement, on suppose que $\epsilon_i \sim F(0, {\sigma_e}^2)$ est indépendant de $\boldsymbol{\beta}$.

Selon le modèle (1.17),

$$\begin{aligned}
\text{E}[Y_i] &= \boldsymbol{X}_i^T\boldsymbol{\gamma}\ , \\
\text{Var}[Y_i] &= \tau\sum_{j=1}^{r}\omega_j G_{ij}^2 + {\sigma_e}^2\ , \\
\text{Covar}[Y_i, Y_l] &= \tau\sum_{j=1}^{r}\omega_j G_{ij}G_{lj}
\end{aligned}$$

pour $i = 1, ..., n$ et $l = 1, ..., n$ avec $i \neq l$.

Pour tester l'association entre Y_i et $\boldsymbol{G_i}$ dans le modèle (1.17), on pose $H_0 : \boldsymbol{\beta} = \boldsymbol{0}_r$. De manière équivalente, on peut tester $H_0 : \tau = 0$ à l'aide d'un test du score, ce

qui ne nécessite pas d'estimer τ. La statistique du score pour la composante de variance τ est donnée par

$$Q_{\text{SKAT}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\gamma})^T \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\gamma}),$$

où $\hat{\gamma}$ est estimée sous l'hypothèse nulle. Il est possible de démontrer que la distribution de Q_{SKAT} sous l'hypothèse nulle est un mélange de χ^2.

CHAPITRE II

COPULES

Un aspect important de notre test d'association est qu'on ne suppose pas la normalité multivariée pour la distribution conjointe des variables dépendantes. Pour arriver à cela, on utilise des copules, qui sont des outils statistiques qui permettent de modéliser la structure de dépendance entre plusieurs variables aléatoires indépendamment des distributions marginales suivies par celles-ci. Plus spécifiquement, une copule est une fonction de répartition multivariée pour laquelle les marginales univariées sont uniformes sur l'intervalle (0,1) (Nelsen, 2006). Dans ce chapitre, on présente quelques définitions et propriétés théoriques de base avant d'illustrer les principales classes de copules, soient les copules archimédiennes et les copules elliptiques.

2.1 Définitions

La fonction de répartition d'une variable aléatoire X est la fonction F_X telle que $\forall x \in \mathbb{R}$,

$$F_X(x) = P[X \leq x].$$

On peut démontrer qu'une fonction de répartition est distribuée suivant une loi Uniforme sur [0,1]. En effet, soient F_X la fonction de répartition strictement croissante d'une variable aléatoire X suivant une distribution quelconque et F_X^{-1} la

fonction inverse de F_X, alors

$$\begin{aligned} P[F_X(x) \leq U] &= P[F_X^{-1}(F_X(x)) \leq F_X^{-1}(U)] \\ &= P[X \leq F_X^{-1}(U)] \\ &= F_X(F_X^{-1}(U)) \\ &= U. \end{aligned}$$

La fonction de répartition conjointe d'un vecteur aléatoire (X, Y) est la fonction H telle que $\forall x, y \in \mathbb{R}^2$,

$$H(x, y) = P[X \leq x, Y \leq y].$$

À partir de la fonction de répartition conjointe, on peut obtenir les fonctions de répartitions marginales par

$$F_X(x) = \lim_{y \to \infty} H(x, y),$$

$$F_Y(y) = \lim_{x \to \infty} H(x, y).$$

De plus, X et Y sont indépendants si et seulement si

$$\begin{aligned} H(x, y) &= P[X \leq x, Y \leq y] \\ &= P[X \leq x]P[Y \leq y] \\ &= F_X(x)F_Y(y), \end{aligned}$$

c'est-à-dire que leur fonction de répartition conjointe est donnée par le produit des marginales.

Afin de relier la fonction de répartition conjointe avec les marginales, on utilise une copule en se basant sur le résultat du Théorème de Sklar.

THÉORÈME 2.1.1 (THÉORÈME DE SKLAR) — *Soit H une fonction de répartition conjointe avec des fonctions marginales F_X et F_Y, alors il existe une copule C telle que $\forall x, y \in \mathbb{R}^2$*

$$H(x, y) = C(F_X(x), F_Y(y)). \tag{2.1}$$

Si F_X et F_Y sont continues, alors la copule C est unique. Dans ce cas, on peut inverser l'équation (2.1) telle que

$$C(u,v) = H(F_X^{-1}(u), F_Y^{-1}(v)) \tag{2.2}$$

où u, v sont uniformément distribués sur $[0,1]^2$.

En dérivant l'équation (2.1), on obtient la fonction de densité conjointe du vecteur (X,Y), soit

$$\begin{aligned} f(x,y) &= \frac{\partial H(x,y)}{\partial x \partial y} \\ &= \frac{\partial C(F_X(x), F_Y(y))}{\partial x \partial y} \\ &= \frac{\partial C(F_X(x), F_Y(y))}{\partial F_X(x) \partial F_Y(y)} \frac{\partial F_X(x)}{\partial x} \frac{\partial F_Y(y)}{\partial y} \\ &= f_X(x)\ f_Y(y)\ c(F_X(x), F_Y(y)) \end{aligned} \tag{2.3}$$

avec $c(F_X(x), F_Y(y))$ définie comme la densité de la copule C.

2.2 Construction de copules par la méthode de l'inverse

L'équation (2.2) est utile pour construire des copules à partir de fonctions de répartition conjointes. Par exemple, soit la fonction exponentielle bivariée de Gumbel donnée par

$$H_\theta(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)}, & x \geq 0, y \geq 0 \\ 0 & \text{ailleurs} \end{cases} \tag{2.4}$$

avec θ un paramètre de dépendance dans [0,1]. Pour trouver la copule de Gumbel bivariée correspondante, on calcule tout d'abord les marginales :

$$\begin{aligned} F_X(x) &= \lim_{y \to \infty} H_\theta(x,y) \\ &= \begin{cases} 1 - e^{x}, & x \geq 0 \\ 0 & \text{ailleurs} \end{cases} \end{aligned}$$

et

$$F_Y(y) = \lim_{x\to\infty} H_\theta(x,y)$$
$$= \begin{cases} 1-e^y, & y \geq 0 \\ 0 & \text{ailleurs} \end{cases}.$$

On isole ensuite x et y

$$x = \ln(1-F_X),$$

$$y = \ln(1-F_Y).$$

D'où, $\forall u, v \in [0,1]$

$$F_X^{-1}(u) = \ln(1-u),$$

$$F_Y^{-1}(v) = \ln(1-v).$$

Enfin, on remplace $F_X^{-1}(u)$ et $F_Y^{-1}(v)$ dans (2.4), ce qui donne

$$C_\theta(u,v) = H(F_X^{-1}(u), F_Y^{-1}(v))$$
$$= u+v-1+(1-u)(1-v)\mathrm{e}^{-\theta\ln(1-u)\ln(1-v)}. \qquad (2.5)$$

La copule bivariée de Gumbel permet de coupler n'importe quelles fonctions marginales continues à une fonction de répartition conjointe exponentielle bivariée selon un paramètre de dépendance θ.

2.3 Copules Archimédiennes

Les copules archimédiennes représentent une classe importante de copules couramment utilisées étant donné la simplicité de leur construction et le fait qu'elles présentent une forme explicite, contrairement aux copules elliptiques. Avant de donner une définition détaillée des copules archimédiennes, on commence par un exemple simple afin d'illuster la manière dont ces copules sont construites.

Soit la copule archimédienne C définie telle que

$$C(u,v) = \frac{uv}{u+v-uv}.$$

On réarrange quelque peu l'équation précédente, d'où

$$\begin{aligned} &\frac{1}{C(u,v)} - 1 = \frac{1}{u} - 1 + \frac{1}{v} - 1 \\ &\Leftrightarrow \varphi(C(u,v)) = \varphi(u) + \varphi(v) \\ &\Leftrightarrow C(u,v) = \varphi^{-1}[\varphi(u) + \varphi(v)] \end{aligned} \tag{2.6}$$

avec $\varphi(t) = \frac{1}{t} - 1$. On peut se baser sur le résultat (2.6) pour définir les copules archimédiennes.

Soit φ une fonction continue strictement décroissante et convexe telle que $\varphi(1) = 0$, et soit φ^{-1} la fonction inverse de φ. Alors, les copules construites sous la forme de (2.6) sont appelées copules archimédiennes. De plus, on définit la fonction φ comme étant le générateur de la copule. Une propriété importante des copules archimédiennes, qui découle directement de (2.6), est qu'elles sont symétriques, c'est-à-dire que

$$C(u,v) = C(v,u)\ \forall u, v \text{ dans } [0,1].$$

Enfin, on présente dans le Tableau 2.1 quelques copules archimédiennes à un paramètre les plus couramment utilisées. Les générateurs et leurs inverses pour chacune des copules sont présentés dans le Tableau 2.2.

2.4 Copules Elliptiques

Les copules elliptiques représentent une autre classe de copules couramment utilisée, notamment en hydrologie et en finance, dû à leur grande flexibilité (Wang et Yan, 2013). Leur construction est basée sur les distributions aléatoires dites elliptiques, dont les plus connues sont la loi normale multivariée et la loi de Student

Tableau 2.1: Copules archimédiennes bivariées

	$C_\theta(u,v)$	$\theta \in$
Clayton	$[\max(u^{-\theta}+v^{-\theta}-1,0)]^{-1/\theta}$	$[-1,\infty)\setminus\{0\}]$
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$[-1,1)$
Gumbel-Hougaard	$\exp(-[(-\ln u)^\theta+(-\ln v)^\theta]^{1/\theta})$	$[1,\infty)$
Frank	$-\frac{1}{\theta}\ln(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1})$	$(-\infty,\infty)\setminus\{0\}$
Joe	$1-[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta]^{1/\theta}$	$[1,\infty)$
Indépendance	uv	

Tableau 2.2: Générateurs des copules archimédiennes du Tableau 2.1

	$\varphi_\theta(t)$	$\varphi_\theta^{-1}(t)$
Clayton	$\frac{1}{\theta}(t^{-\theta}-1)$	$(1+\theta t)^{-1/\theta}$
Ali-Mikhail-Haq	$\ln\frac{1-\theta(1-t)}{t}$	$\frac{1-\theta}{e^t-\theta}$
Gumbel-Hougaard	$(-\ln t)^\theta$	$\exp(-t^{1/\theta})$
Frank	$-\ln\frac{e^{-\theta t}-1}{e^\theta-1}$	$-\frac{1}{\theta}\ln(1+e^{-t}(e^{-\theta}-1))$
Joe	$-\ln[1-(1-t)^\theta]$	$1-(1-e^{-t})^{1/\theta}$
Indépendance	$-\ln t$	e^{-t}

multivariée. On définit tout d'abord le concept de distribution aléatoire elliptique, avant de présenter les copules elliptiques associées.

Soit un vecteur aléatoire $\boldsymbol{X}$ de dimension p suivant une loi elliptique, de moyenne $\boldsymbol{\mu}$ et de variance-covariance $\boldsymbol{\Sigma}$. Alors, la fonction de densité de $\boldsymbol{X}$ est donnée par

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

avec g définie comme la fonction génératrice de densité. Lorsque $g(t) = (2\pi)^{-n/2} e^{-t/2}$, $\boldsymbol{X}$ suit une loi normale de dimension p. En faisant varier g, il est possible d'obtenir des lois pour lesquelles les queues de distribution sont plus ou moins longues que pour la loi normale. En se basant sur l'équation (2.2), on obtient la copule elliptique implicite pour $\boldsymbol{X}$, soit

$$C(u_1, ..., u_p) = H\{G_1^{-1}(u_1), ..., G_p^{-1}(u_p)\}, \qquad u_i \in (0,1)\ , i = 1, ..., p,$$

où H est la fonction de répartition conjointe de $\boldsymbol{X}$ et G_i^{-1} est la fonction de répartition univariée inverse de X_i.

2.5 Dépendance

Soient (X_1, Y_1) et (X_2, Y_2) deux vecteurs aléatoires bivariés indépendants et identiquement distribués selon une fonction de répartiton conjointe H, alors on définit le tau de Kendall comme

$$\tau_{XY} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \qquad (2.7)$$

En d'autres mots, le tau de Kendall mesure la différence entre la probabilité de concordance et la probabilité de discordance pour des paires d'observations . De plus, la valeur de τ_{XY} dans (2.7) est toujours comprise entre -1 et 1, étant donné qu'il s'agit d'une différence entre deux probabilités complémentaires. On peut démontrer la relation importante entre les copules et le tau de Kendall par le Théorème 2.5.1.

THÉORÈME 2.5.1 — *Soient X et Y deux variables aléatoires continues liées par une copule C. Alors, le tau de Kendall pour X et Y est donné par*

$$\begin{aligned} \tau_{XY} &= 4 \int \int C(u,v)\ dC(u,v) - 1 \\ &= 4\ E[C(U,V)] - 1, \end{aligned} \tag{2.8}$$

avec U et V des variables aléatoires uniformes sur [0,1] dont la fonction de répartition conjointe est C.

L'équation (2.8) est très importante, car elle permet une transformation du tau de Kendall vers le ou les paramètres de dépendance de n'importe quelle copule. On peut ainsi comparer la dépendance à travers différentes copules en utilisant un paramètre de concordance dont le support est le même pour chaque copule.

Dans le cas des copules elliptiques, on peut utiliser le résultat du Théorème 2.5.2 (Fang et Fang, 2002).

THÉORÈME 2.5.2 — *Soient X et Y deux variables aléatoires continues liées par une copule elliptique. Alors, le tau de Kendall pour X et Y est donné par*

$$\tau_{XY} = \frac{2}{\pi} arcsin(\rho), \tag{2.9}$$

où ρ est le coefficient de corrélation de Pearson entre X et Y.

Dans le cas des copules archimédiennes, on peut calculer le tau de Kendall directement à partir du générateur de la copule, ce qui est en général plus simple que d'évaluer l'intégrale double en (2.8).

THÉORÈME 2.5.3 — *Soient X et Y deux variables aléatoires continues liées par une copule archimédienne C générée par φ. Alors le tau de Kendall pour X et Y est donné par*

$$\tau_{XY} = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt.$$

CHAPITRE III

MÉTHODOLOGIE

Dans ce chapitre, on introduit tout d'abord le modèle linéaire mixte du test de MURAT avant de présenter la méthodologie derrière notre test d'association CBMRV.

3.1 Test d'association multivarié basé sur un modèle linéaire mixte (MURAT)

Supposons qu'on observe 2 traits pour N individus indépendants, $\boldsymbol{Y}_i = (y_{i1}, y_{i2})^T$, $i = 1, ..., N$, m covariables, $\boldsymbol{X_i} = (x_{i1}, x_{i2}, ..., x_{im})^T$, et un ensemble de SNPs, $\boldsymbol{G_i} = (g_{i1}, g_{i2}, ..., g_{ir})^T$ contenant r variants bialléliques codés $0, 1, 2$, représentant le nombre d'allèles mineurs. On s'intéresse à l'association entre $\boldsymbol{Y_i}$ et $\boldsymbol{G_i}$ en supposant le modèle linéaire mixte (Sun *et al.*, 2016) tel que

$$\boldsymbol{Y_i} = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_i^T \boldsymbol{\gamma}_1 \\ \boldsymbol{X}_i^T \boldsymbol{\gamma}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{G}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{G}_i^T \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \text{ pour } i = 1, ..., N, \tag{3.1}$$

où pour le k^e phénotype, $k = 1, 2$, $\boldsymbol{\gamma_k} = (\gamma_{k1}, ..., \gamma_{km})^T$, et $\boldsymbol{\beta_k} = (\beta_{k1}, ..., \beta_{kr})^T$ sont les coefficients de régression pour $\boldsymbol{X_i}$ et $\boldsymbol{G_i}$ respectivement. De plus, on suppose que $\boldsymbol{\gamma_k}$ est un vecteur d'effets fixes tandis que $\boldsymbol{\beta_k} \sim N(\boldsymbol{0}_r, \tau \boldsymbol{W})$ est un vecteur d'effets aléatoires, avec τ une composante de variance inconnue et $\boldsymbol{W} = \text{diag}(w_1, ..., w_r)$ une matrice de poids telle que w_j est le poids associé au j^e

variant. Finalement, on suppose que $\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N_2(\mathbf{0}_2, \boldsymbol{\Sigma}_e)$ et que

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2r}(\mathbf{0}_{2r}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \tau \begin{bmatrix} \boldsymbol{W} & \rho\boldsymbol{W} \\ \rho\boldsymbol{W} & \boldsymbol{W} \end{bmatrix}),$$

où ρ est un paramètre inconnu qui sert à capturer l'effet pléiotropique de la région.

Selon le modèle (3.1), la distribution marginale de $\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}$ est égale à

$$f(\boldsymbol{y}_i) = \int f(\boldsymbol{y}_i, \boldsymbol{\beta})\, d\boldsymbol{\beta}$$

$$= \int f(\boldsymbol{y}_i|\boldsymbol{\beta})\, f(\boldsymbol{\beta})\, d\boldsymbol{\beta}$$

qui est une normale bivariée étant donné que $\boldsymbol{\epsilon}$ et $\boldsymbol{\beta}$ suivent des lois normales. Ainsi, il est possible de montrer que

$$f(\boldsymbol{y}_i) = N_2 \left[\begin{pmatrix} \boldsymbol{X}_i^T\boldsymbol{\gamma}_1 \\ \boldsymbol{X}_i^T\boldsymbol{\gamma}_2 \end{pmatrix}, \tau R \otimes \boldsymbol{G}_i^T\boldsymbol{W}\boldsymbol{G}_i + \boldsymbol{\Sigma}_e \right],$$

où $R = (1-\rho)\boldsymbol{I}_2 + \rho\mathbf{1}_2\mathbf{1}_2^T$ et $\mathbf{1}_2 = (1,1)^T$. Donc, la vraisemblance marginale est donnée comme suit

$$L(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \tau, \boldsymbol{\Sigma}_\epsilon|\boldsymbol{y}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i)$$

$$= \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{Y} - \begin{pmatrix} \boldsymbol{X}\boldsymbol{\gamma}_1 \\ \boldsymbol{X}\boldsymbol{\gamma}_2 \end{pmatrix})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \begin{pmatrix} \boldsymbol{X}\boldsymbol{\gamma}_1 \\ \boldsymbol{X}\boldsymbol{\gamma}_2 \end{pmatrix})\},$$

avec $\boldsymbol{\Sigma} = \tau R \otimes \boldsymbol{G}^T\boldsymbol{W}\boldsymbol{G} + \boldsymbol{\Sigma}_e \otimes \boldsymbol{I}_n$ et $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_N \end{pmatrix}$.

Pour tester l'association entre $\boldsymbol{Y}_i$ et $\boldsymbol{G}_i$ dans le modèle (3.1), on pose $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$. De manière équivalente, on peut tester $H_0 : \tau = 0$ à l'aide d'un test de

score, ce qui ne nécessite pas d'estimer τ. Pour un ρ fixé, la statistique de score pour la composante de variance τ est donnée par

$$S_\rho = (\boldsymbol{Y} - \begin{pmatrix} \boldsymbol{X}\hat{\gamma}_1 \\ \boldsymbol{X}\hat{\gamma}_2 \end{pmatrix})^T \left(\hat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{G}^* [R \otimes W] \boldsymbol{G}^{*T} \hat{\boldsymbol{\Sigma}}_e^{-1} \right) (\boldsymbol{Y} - \begin{pmatrix} \boldsymbol{X}\hat{\gamma}_1 \\ \boldsymbol{X}\hat{\gamma}_2 \end{pmatrix}),$$

où $\hat{\gamma}_1$, $\hat{\gamma}_2$ et $\hat{\boldsymbol{\Sigma}}_e^{-1}$ sont estimés sous l'hypothèse nulle. Toujours pour ρ fixé, il est possible de démontrer que la distribution de S_ρ sous l'hypothèse nulle est un mélange de χ^2.

En pratique, on travaille souvent avec des traits qui ne sont pas distribués selon une loi normale. Dans ces situations, il n'est donc pas possible de supposer la normalité multivariée du vecteur d'erreurs $\boldsymbol{\epsilon}$, étant donné que cela impliquerait automatiquement une distribution marginale normale pour chaque trait composant le phénotype. De plus la normalité multivariée pour le phénotype $\boldsymbol{Y_i}$ implique une dépendance linéaire entre les deux traits Y_{i1} et Y_{i2}, ce qui n'est pas toujours le cas en réalité. Pour ces raisons, nous présentons dans la prochaine section un nouveau test d'association multivarié, CBM-RV, pour lequel on relaxe l'hypothèse de normalité pour la distribution de la variable dépendante à l'aide de copules.

3.2 Test d'association CBM-RV

On suppose que les distributions marginales de Y_1 et Y_2 proviennent des familles exponentielles et que $\boldsymbol{\beta}$ suit une loi quelconque telle que

$$\boldsymbol{\beta} \sim F(\mathbf{0}_{2r}, \boldsymbol{\Sigma}_\beta = \tau \begin{bmatrix} \boldsymbol{W} & \rho\boldsymbol{W} \\ \rho\boldsymbol{W} & \boldsymbol{W} \end{bmatrix}). \tag{3.2}$$

On obtient, pour $j = 1, 2$, un modèle linéaire généralisé mixte pour lequel

$$\begin{aligned} \mathrm{E}[Y_{ij}|\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{\beta}_j] &= \mu_{ij} = g_j^{-1}\left(\boldsymbol{X}_i^T\boldsymbol{\gamma}_j + \boldsymbol{G}_i^T\boldsymbol{\beta}_j\right) \\ \mathrm{Var}[Y_{ij}|\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{\beta}_j] &= \phi_j \cdot \nu_j(\mu_{ij}), \end{aligned} \tag{3.3}$$

où g est la fonction de lien, ϕ est le paramètre de dispersion et ν la fonction variance. Par le théorème de Sklar, on modélise la dépendance entre les deux traits par une copule C_α, qui dépend d'un paramètre de dépendance α. Donc, par l'équation (2.1), la fonction de répartition conjointe conditionnelle est donnée par

$$H(y_{i1}, y_{i2}|\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{\beta}) = C_\alpha(F_1(y_{i1}|\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{\beta}), F_2(y_{i2}|\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{\beta})). \tag{3.4}$$

La fonction de densité conjointe conditionnelle est obtenue en dérivant (3.4), telle que démontrée dans l'équation (2.3). Ainsi, en omettant les termes $\boldsymbol{X}_i$ et $\boldsymbol{G}_i$ pour abréger la notation, on trouve

$$f(y_{i1}, y_{i2}|\boldsymbol{\beta}) = f_1(y_{i1}|\boldsymbol{\beta}) f_2(y_{i2}|\boldsymbol{\beta}) c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}), F_2(y_{i2}|\boldsymbol{\beta})). \tag{3.5}$$

Soit $\boldsymbol{\theta} = (\alpha, \tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \phi_1, \phi_2)^T$ le vecteur contenant les paramètres du modèle en (3.3), la distribution conditionnelle des phénotypes est

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_{i1}, y_{i2}|\boldsymbol{\beta}),$$

puisqu'en conditionnant sur $\boldsymbol{\beta}$, les sujets sont maintenant indépendants. On insère la densité obtenue en (3.5), ce qui donne

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} f_1(y_{i1}|\boldsymbol{\beta}) f_2(y_{i2}|\boldsymbol{\beta}) \cdot c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}); F_2(y_{i2}|\boldsymbol{\beta})). \tag{3.6}$$

La vraisemblance marginale pour les paramètres du modèle est donnée en intégrant la fonction de densité conjointe sur $\boldsymbol{\beta}$, d'où

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int_{\boldsymbol{\beta}} f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})\, d\boldsymbol{\beta} \\ &= \int_{\boldsymbol{\beta}} \prod_{i=1}^{n} f_1(y_{i1}|\boldsymbol{\beta}) f_2(y_{i2}|\boldsymbol{\beta}) \cdot c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}); F_2(y_{i2}|\boldsymbol{\beta})) f(\boldsymbol{\beta})\, d\boldsymbol{\beta}. \end{aligned} \tag{3.7}$$

On définit la vraisemblance conditionnelle du modèle telle que

$$L(\boldsymbol{\theta}|\boldsymbol{\beta}) = f(\boldsymbol{y}|\boldsymbol{\beta}). \tag{3.8}$$

On insère (3.8) dans l'équation (3.7), d'où

$$L(\boldsymbol{\theta}) = \int_{\boldsymbol{\beta}} L(\boldsymbol{\theta}|\boldsymbol{\beta}) f(\boldsymbol{\beta})\, d\boldsymbol{\beta}. \tag{3.9}$$

Le test du score pour tester $H_0 : \tau = 0$ ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}_r$) se base sur le score

$$U(\tau) = \frac{\partial}{\partial \tau} \log L(\boldsymbol{\theta}). \tag{3.10}$$

Le problème avec le calcul direct de ce score est l'évaluation de l'intégrale en dimension $\mathbb{R}^{2r}$ dans l'équation (3.9). Pour remédier au problème computationnel, on opte pour une approximation de cette intégrale en utilisant des techniques du développement de Taylor de $L(\boldsymbol{\theta}|\boldsymbol{\beta})$ au voisinage de $\boldsymbol{\beta} = \mathbf{0}_{2r}$ (Lin, 1997).

Ainsi, nous avons

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{\beta}) &= \exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\} \\ &\approx \exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\}_{|\boldsymbol{\beta}=\mathbf{0}} + \frac{\partial}{\partial \boldsymbol{\beta}} \left[\exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\}\right]^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \\ &\qquad + \frac{1}{2}\boldsymbol{\beta}^T \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} [\exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\}]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta}. \end{aligned} \tag{3.11}$$

En développant les dérivées, nous pouvons écrire

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{\beta}) &\approx L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} + \left[\exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\} \cdot \frac{\partial}{\partial \beta} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right]^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \\ &\qquad + \frac{1}{2}\boldsymbol{\beta}^T \frac{\partial}{\boldsymbol{\beta}} \left[\exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\} \cdot \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \\ &= L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} + L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \\ &\qquad + \frac{1}{2}\boldsymbol{\beta}^T \left[\exp\{l(\boldsymbol{\theta}|\boldsymbol{\beta})\} \left(\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right)\right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \\ &= L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \left[1 + \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \right. \\ &\qquad \left. + \frac{1}{2}\boldsymbol{\beta}^T \left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta}\right]. \end{aligned} \tag{3.12}$$

L'équation (3.12) permet de réécrire la vraisemblance marginale de l'équation (3.9) comme suit

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int_{\boldsymbol{\beta}} L(\boldsymbol{\theta}|\boldsymbol{\beta}) f(\boldsymbol{\beta})\, d\boldsymbol{\beta} \\ &\approx L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \int_{\boldsymbol{\beta}} \left[1 + \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \right. \\ &\qquad \left. + \frac{1}{2} \boldsymbol{\beta}^T \left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \boldsymbol{\beta} \right] f(\boldsymbol{\beta})\, d\boldsymbol{\beta}. \end{aligned}$$

Par définition de la fonction de densité, on sait que $\int_{\boldsymbol{\beta}} f(\boldsymbol{\beta})\, d\boldsymbol{\beta} = 1$ et $\int_{\boldsymbol{\beta}} \boldsymbol{\beta} f(\boldsymbol{\beta})\, d\boldsymbol{\beta} = \mathrm{E}[\boldsymbol{\beta}]$. De plus, l'espérance du terme quadratique étant donné par l'équation (1.3), on approxime l'intégrale précédente comme suit

$$\begin{aligned} L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \Big[1 &+ \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \mathrm{E}[\boldsymbol{\beta}] \\ &+ \frac{1}{2} \mathrm{tr} \Big\{ \Big[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \Big]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \mathrm{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T] \Big\} \Big]. \end{aligned}$$

Puisque $\boldsymbol{\beta}$ est centré autour de $\mathbf{0}_{2r}$, on a

$$\begin{aligned} L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \Big[1 &+ \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \mathbf{0}_{2r} \\ &+ \frac{1}{2} \mathrm{tr} \Big\{ \Big[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \Big]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \mathrm{Var}[\boldsymbol{\beta}] \Big\} \Big]. \end{aligned}$$

Enfin, le développement de Taylor de $L(\boldsymbol{\theta}|\boldsymbol{\beta})$ au voisinage de $\boldsymbol{\beta} = \mathbf{0}_{2r}$ a permis d'approximer la vraisemblance marginale par

$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}|\boldsymbol{\beta})_{|\boldsymbol{\beta}=\mathbf{0}} \left[1 + \frac{1}{2} \mathrm{tr} \left\{ \left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \tau \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \right\} \right]. \tag{3.13}$$

À l'aide de (3.13), on évalue le score de l'équation (3.10), soit

$$\begin{aligned} U(\tau) &= \frac{\partial}{\partial \tau} l(\boldsymbol{\theta}) \\ &\approx \frac{\partial}{\partial \tau} \log \left(1 + \frac{\tau}{2} \mathrm{tr} \left\{ \left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \right\} \right). \end{aligned}$$

Le développement en série de Taylor autour du point $x = 0$ de la fonction $\log(1+x)$ est donné par $\log(1+x) \approx x$. Étant donnée que $\tau = 0$ sous H_0, on trouve que

$$\begin{aligned} U(\tau) &\approx \frac{\partial}{\partial \tau}\left[\frac{\tau}{2}\mathrm{tr}\left\{\left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right]_{|\boldsymbol{\beta}=0} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\right\}\right] \\ &= \frac{1}{2}\mathrm{tr}\left\{\left[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T + \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})\right]_{|\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\right\} \\ &= \frac{1}{2}\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) \cdot \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T|_{\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} + \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\right\}. \end{aligned}$$

Enfin, on applique les propriétés de la trace données en (1.1) et (1.2), d'où

$$\begin{aligned} U(\tau) &\approx \frac{1}{2}\mathrm{tr}\{\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T|_{\boldsymbol{\beta}=\mathbf{0}} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}}\} + \frac{1}{2}\mathrm{tr}\{\frac{\partial^2}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}. \\ &= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T|_{\boldsymbol{\beta}=\mathbf{0}} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} + \frac{1}{2}\mathrm{tr}\{\frac{\partial^2}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\}. \end{aligned} \tag{3.14}$$

Afin de calculer le score obtenu en (3.14), il faut trouver la forme explicite pour la log-vraisemblance conditionnelle $l(\boldsymbol{\theta}|\boldsymbol{\beta})$ ainsi que ses deux premières dérivées évaluées en $\boldsymbol{\beta} = \mathbf{0}$. À partir de (3.6) et (3.8), on trouve

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{\beta}) &= \log L(\boldsymbol{\theta}|\boldsymbol{\beta}) \\ &= \log \prod_{i=1}^{n} f_1(y_{i1}|\boldsymbol{\beta}) f_2(y_{i2}|\boldsymbol{\beta}) \cdot c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}); F_2(y_{i2}|\boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} [\log f_1(y_{i1}|\boldsymbol{\beta}) + \log f_2(y_{i2}|\boldsymbol{\beta}) + \log c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}); F_2(y_{i2}|\boldsymbol{\beta}))]. \end{aligned} \tag{3.15}$$

Les dérivées première et seconde de $l(\boldsymbol{\theta}|\boldsymbol{\beta})$ sont explicitées dans la Proposition 1, avec les détails des calculs présentés dans l'appendice B. Dans la dernière section du présent chapitre, on trouve la distribution du score $U(\tau)$ sous H_0.

Proposition 1 Soit $l(\boldsymbol{\theta}|\boldsymbol{\beta})$ la log-vraisemblance conditionnelle du modèle, telle que donnée en (3.15), alors les dérivées première et seconde en $\boldsymbol{\beta}$ sont données par

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \begin{bmatrix} \frac{1}{\phi} \frac{y_{i1}-\mu_{i1}}{v(\mu_{i1}) \frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} \\ \frac{1}{\phi} \frac{y_{i2}-\mu_{i2}}{v(\mu_{i2}) \frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i.$$

et

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial}{\partial \mu_{i1}} [\frac{1}{\phi} \frac{y_{i1}-\mu_{i1}}{v(\mu_{i1}) g'_{\mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}] \frac{1}{g'_{\mu_{i1}}} & \frac{\partial}{\partial \mu_{i1}} [\frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}] \frac{1}{g'_{\mu_{i1}}} \\ \frac{\partial}{\partial \mu_{i2}} [\frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}] \frac{1}{g'_{\mu_{i2}}} & \frac{\partial}{\partial \mu_{i2}} [\frac{1}{\phi} \frac{y_{i2}-\mu_{i2}}{v(\mu_{i2}) g'_{\mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}] \frac{1}{g'_{\mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i \boldsymbol{G}_i^T.$$

3.3 Distribution de la statistique du test CBM-RV

Soit Q_ρ le score défini dans l'équation (3.14), notons

$$Q_1 = \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T|_{\boldsymbol{\beta}=\mathbf{0}} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} \tag{3.16}$$

et

$$Q_2 = \frac{1}{2} \mathrm{tr}\{ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \} \tag{3.17}$$

tel que $Q_\rho \approx Q_1 + Q_2$. Il est possible de démontrer que la distribution théorique de Q_ρ, pour ρ fixé, est un mélange de χ^2_1. Tout d'abord, on peut réécrire l'équation (3.16) telle que

$$Q_1 = \frac{1}{2} \boldsymbol{L}^T \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{L} \qquad \text{, avec } \boldsymbol{L} = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta})^T|_{\boldsymbol{\beta}=\mathbf{0}}$$

$$= \frac{1}{2} (\boldsymbol{\Sigma}_{\boldsymbol{L}}^{-1/2} \boldsymbol{L})^T \boldsymbol{\Sigma}_{\boldsymbol{L}}^{1/2} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{L}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{L}}^{-1/2} \boldsymbol{L}$$

où $\boldsymbol{\Sigma_L}$ est la matrice de variance-covariance du score $\boldsymbol{L}$, qui suit une loi normale centrée sous $H_0 : \boldsymbol{\beta} = \mathbf{0}_{2r}$. Soit le score centré et réduit $\boldsymbol{Z} = \boldsymbol{\Sigma_L^{-1/2}} \boldsymbol{L}$, alors on a

$$Q_1 = \frac{1}{2} \boldsymbol{Z}^T \boldsymbol{\Sigma_L^{1/2}} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{\Sigma_L^{1/2}} \boldsymbol{Z}.$$

En utilisant la décomposition spectrale de l'équation (1.5) sur la matrice $\boldsymbol{\Sigma_L^{1/2}} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{\Sigma_L^{1/2}}$, on obtient

$$\begin{aligned} Q_1 &= \frac{1}{2} \boldsymbol{Z}^T (\sum_{i=1}^{2r} \lambda_i \boldsymbol{u_i} \boldsymbol{u_i^T}) \boldsymbol{Z} \\ &= \frac{1}{2} \sum_{i=1}^{2r} \lambda_i (\boldsymbol{Z}^T \boldsymbol{u_i})(\boldsymbol{u_i^T} \boldsymbol{Z}). \end{aligned}$$

Or, $\boldsymbol{Z^T u_i} = \sum_{j=1}^{2r} u_{ij} Z_j$ est une somme de variables indépendantes normales centrées et réduites puisque $\boldsymbol{Z} \sim N_{2r}(\mathbf{0}_{2r}, \boldsymbol{I}_{2r})$. De plus, $\boldsymbol{u_i}$ est un vecteur propre normalisé dont la norme est égale à un. Il est donc facile de montrer que $\boldsymbol{Z^T u_i} \sim N(0,1)$. Par conséquent, on trouve

$$\begin{aligned} Q_1 &= \frac{1}{2} \sum_{i=1}^{2r} \lambda_i (\boldsymbol{Z}^T \boldsymbol{u_i})^2 \\ &= \frac{1}{2} \sum_{i=1}^{2r} \lambda_i \chi_1^2. \end{aligned} \tag{3.18}$$

Ensuite, on réécrit l'équation (3.17) telle que

$$Q_2 = \frac{1}{2} \mathrm{tr}\{\boldsymbol{B} \cdot \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\} \qquad \text{, avec } \boldsymbol{B} = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}}$$

$$= \frac{1}{2} \mathrm{tr}\{\boldsymbol{B}^{1/2} \begin{bmatrix} \boldsymbol{W} & \rho \boldsymbol{W} \\ \rho \boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{B}^{1/2}\} \qquad \text{par (1.2).}$$

En utilisant la propriété de l'équation (1.4), on obtient finalement que

$$Q_2 = \frac{1}{2} \sum_{i=1}^{2r} \lambda_i^*, \tag{3.19}$$

où les λ_i^*, $i = 1, ..., 2r$, sont les valeurs propres de $\boldsymbol{B}^{1/2} \begin{bmatrix} \boldsymbol{W} & \rho\boldsymbol{W} \\ \rho\boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{B}^{1/2}$.

On combine les équations (3.18) et (3.19) , d'où

$$Q_\rho \approx \frac{1}{2}\sum_{i=1}^{2r} \lambda_i \chi_1^2 + \frac{1}{2}\sum_{i=1}^{2r} \lambda_i^*, \tag{3.20}$$

où les λ_i, $i = 1, ..., 2r$, sont les valeurs propres de $\boldsymbol{\Sigma}_{\boldsymbol{L}}^{1/2} \begin{bmatrix} \boldsymbol{W} & \rho\boldsymbol{W} \\ \rho\boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{L}}^{1/2}$.

Lorsque les paramètres de nuisance du modèle sont connus, soient α, γ_1, γ_2, ϕ_1, ϕ_2 et ρ, on sait par (1.15) que $\boldsymbol{\Sigma_L} = \boldsymbol{I}(\boldsymbol{\beta})$. En pratique, on remplace $\boldsymbol{I}(\boldsymbol{\beta})$ par la matrice d'information observée de Fisher, soit

$$\boldsymbol{I}_{obs}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}.$$

Puisqu'on évalue $\boldsymbol{I}_{obs}(\boldsymbol{\beta})$ sous H_0, on obtient que $\boldsymbol{\Sigma_L} = -\boldsymbol{B}$. Donc, l'équation (3.20) devient

$$Q_\rho \approx \frac{1}{2}\sum_{i=1}^{2r} \lambda_i(\chi_1^2 - 1). \tag{3.21}$$

Nous rencontrons deux problèmes majeurs lors du calcul de la distribution de Q_ρ en (3.21). Tout d'abord, la valeur de ρ est généralement inconnue. Pour construire une statistique de test pour ρ inconnu, on pourrait appliquer une approche qui s'adapte aux données en sélectionnant, parmi une grille de valeurs sur $[0, 1]$, la valeur optimale de ρ qui maximize la puissance du test (Lee *et al.*, 2012). Dans le cadre de ce mémoire, nous contournons cette difficulté en supposant $\rho = 0$ dans l'équation (3.2). En d'autres mots, on suppose qu'il n'y a pas de corrélation entre les effets d'un même variant sur différents phénotypes. Il est toutefois important de mentionner qu'en pratique, lorsque les deux phénotypes sont fortement corrélés, cette hypothèse n'est pas réaliste et elle peut entraîner une perte de puissance importante.

Deuxièmement, même pour un ρ fixé, on ne connaît jamais, en pratique, la vraie valeur des autres paramètres de nuisance du modèle, soient α, γ_1, γ_2, ϕ_1 et ϕ_2. Si on utilise les estimateurs de maximum de vraisemblance de ces paramètres, l'équation (3.21) ne permet pas d'approximer adéquatement la distribution du score Q_ρ sous H_0, tel que démontré dans les simulations du Chapitre 4. Pour remédier à ce problème, on doit tenir compte des paramètres de nuisance du modèle dans le calcul de la matrice d'information de Fisher observée. Ainsi, la distribution du score Q_ρ sous H_0, correctement ajustée pour les paramètres de nuisance, est donnée en (3.20), avec $\boldsymbol{\Sigma_L}$ définie dans l'équation (1.16).

CHAPITRE IV

ÉTUDE DE SIMULATION

Dans ce chapitre, on vérifie les propriétés du test CBM-RV à l'aide de différents scénarios de simulation. Le premier consiste à simuler la distribution du score sous l'hypothèse nulle de non-association entre $\boldsymbol{Y}$ et $\boldsymbol{G}$ pour différentes lois marginales. Le deuxième scénario de simulations porte sur les erreurs de spécification du modèle, c'est-à-dire sur l'inflation de l'erreur de type I lorsqu'on spécifie soit la mauvaise copule, soit la mauvaise distribution pour les marges. Finalement, le troisième scénario porte sur la puissance du test CBM-RV.

4.1 Algorithme de simulation

Afin de simuler des génotypes, on utilise la base de données du 1000 Genomes Project, qui est une des plus grandes ressources publiques de données génotypiques disponibles, avec à ce jour un total de 2504 sujets provenant de différentes populations ciblées (IGSR, 2017). Plus spécifiquement, on sélectionne des variants génétiques situés sur le gène BRCA1, pour lequel plusieurs mutations connues sont associées à un risque plus élevé de développer les cancers du sein, des ovaires et de la prostate (National Library of Medicine, 2015). Pour l'ensemble des scénarios, on sélectionne les 503 sujets avec une ascendance génétique européenne afin d'éviter les structures de populations dans l'échantillon, ce qui pourrait potentiellement créer de l'inflation sur l'erreur de type I. Enfin, l'algorithme de simulation pour

les phénotypes est le suivant :

1. Simuler N observations d'une covariable X_1 suivant une loi de Bernoulli $(p = 0.5)$ et N observations d'une covariable X_2 suivant une loi Normale $(0, 1)$.

2. Fixer une ordonnée à l'origine et des effets fixes pour les covariables X_1 et X_2, tels que

$$\gamma_1 = (0.58, 1.58, 0.82)^T$$
$$\gamma_2 = (1.77, 1.88, 0.09)^T.$$

3. Sélectionner aléatoirement r variants rares situés sur le gène BRCA1.

4. Fixer les poids w_j, $j = 1, ..., r$ pour les r variants génétiques tels que $\boldsymbol{W} = \text{diag}(w_1, ..., w_r)$.

5. Simuler le vecteur d'effets aléatoires $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2r}\left(\mathbf{0}_{2r}, \tau \begin{bmatrix} \boldsymbol{W} & \rho\boldsymbol{W} \\ \rho\boldsymbol{W} & \boldsymbol{W} \end{bmatrix}\right)$, avec $\tau > 0$ sous H_A. Sous H_0, poser $\tau = 0$.

6. Simuler N couples de quantiles $(U, V) \in [0, 1]^2$ à partir de la copule gaussienne, sachant que

$$u_i = \Phi(z_{i1}),$$
$$v_i = \Phi(z_{i2}),$$

avec $\boldsymbol{Z_i} = (z_{i1}, z_{i2})^T \sim N_2\left(\mathbf{0}_2, \begin{bmatrix} 1 & \rho_e \\ \rho_e & 1 \end{bmatrix}\right)$.

7. Fixer les paramètres de dispersion $\phi_1 = \phi_2 = 1$ et simuler N observations du couple de phénotypes (Y_1, Y_2) telles que

$$y_{i1} = F_1^{-1}(u_i|\mu_{i1}, \phi_1, \tau),$$
$$y_{i2} = F_2^{-1}(v_i|\mu_{i2}, \phi_2, \tau),$$

avec $i = 1, ..., N$ et sachant que

$$\mu_{ik} = g_k^{-1}(\boldsymbol{X}_i^T \boldsymbol{\gamma}_k + \boldsymbol{G}_i^T \boldsymbol{\beta}_k),$$

où $F_k^{-1}(\cdot)$ et $g_k^{-1}(\cdot)$ sont respectivement la fonction de répartition inverse et la fonction de lien inverse pour le trait Y_k, $k = 1, 2$.

8. Répéter les étapes 1 à 7 pour obtenir m échantillons indépendants.

Pour les scénarios 1 et 2, nous allons fixer le nombre de réplications à 10 000, tandis qu'il est fixé à 5000 pour le scénario 3. Pour tous les scénarios, on va considérer une région génétique composée de $r = 30$ variants rares, sélectionnés aléatoirement à chaque itération parmi les 122 variants rares du gène BRCA1. En sélectionnant un nouvel échantillon aléatoire de variants rares pour chaque itération, on cherche à évaluer l'impact de la distribution des fréquences alléliques sur la distribution de notre statistique de test. Pour tous les scénarios, on pose $w_j = 1$, pour $j = 1, ..., 15$. Enfin, la corrélation ρ_e entre les erreurs du modèle est fixé à 0.5 pour les scénarios 1 et 2.

4.2 Scénario 1 : Distribution du score du test CBM-RV sous H_0

La distribution théorique du score pour le test CBM-RV sous H_0 étant démontrée dans la section 3.3, on peut comparer, à l'aide de diagrammes quantile-quantile, les scores obtenus pour 10 000 réplications avec les scores théoriques attendus. Tout d'abord, on démontre que la distribution simplifiée de l'équation (3.21) n'est effectivement pas valable lorsqu'on doit estimer les paramètres de nuisance du modèle sous l'hypothèse nulle. Dans la section suivante, on démontre que l'équation (3.20) approxime adéquatement la distribution du score du test CBM-RV sous H_0 lorsque les paramètres de nuisance sont estimés.

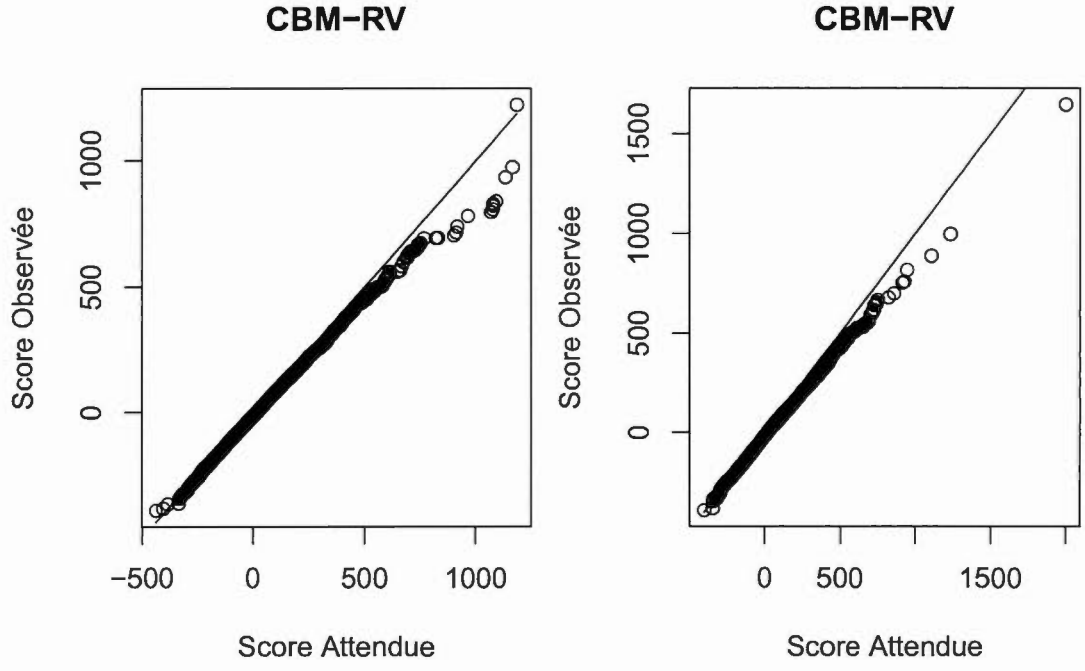
4.2.1 Distribution du score sous H_0 sans correction

On simule deux traits corrélés de lois marginales Normales ou Gamma en suivant l'algorithme présenté dans la section 4.1. Ensuite, on estime les paramètres de nuisance du modèle, toujours sous $H_0 : \tau = 0$, à l'aide de la méthode du maximum de vraisemblance. Enfin, on suppose que la statistique du score suit approximativement la distribution théorique en (3.21). Le Tableau 4.1 présente les valeurs de l'estimation empirique de l'erreur de type I du test CBM-RV pour différentes valeurs du seuil α lorsque les phénotypes sont respectivement de lois marginales Normales et Gamma. Les diagrammes quantile-quantile pour la statistique du test sous H_0 ainsi que les p-valeurs associées sont présentés dans les Figures 4.1 et 4.2.

Tableau 4.1: Estimation de l'erreur de type I du test du score CBM-RV sans correction pour l'estimation des paramètres de nuisance.

	$H_0 : \tau = 0$		
α	0.01	0.05	0.1
Lois Normales	0.0056	0.0364	0.0775
Lois Gamma	0.0049	0.0299	0.0677

En observant les diagrammes de la Figure 4.2, on constate qu'il y a un problème avec la distribution des p-valeurs du score. En effet, il semble qu'il y ait une surestimation de la variance de la statistique du score, d'où la sous-dispersion observée par rapport à la distribution théorique attendue. Cette sous-dispersion peut affecter de façon importante la puissance du test du score sous l'alternative. Pour corriger cette estimation biaisée dans la variance de la statistique, il faut tenir compte de la variation additionnelle introduite par les estimateurs des paramètres de nuisance du modèle dans le calcul de la matrice d'information de Fisher, telle que définie dans l'équation (1.16).

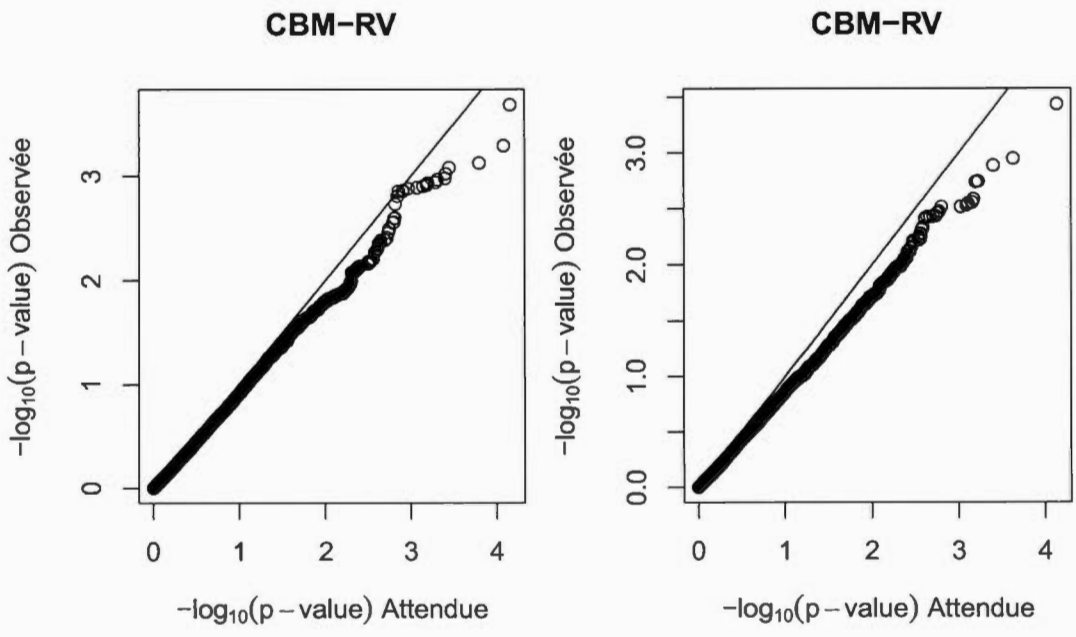


(a) (b)

Figure 4.1: Diagrammes quantile-quantile de la statistique du test CBM-RV pour 10 000 simulations sous H_0 sans correction pour l'estimation des paramètres de nuisance et pour Y_1 et Y_2 de lois marginales : (a) Normales (b) Gamma.

4.2.2 Distribution du score sous H_0 avec correction

On reprend le scénario de simulation précédent, en appliquant la correction pour l'estimation de la variance du score. En d'autres mots, on suppose que la statistique de score suit plutôt la distribution théorique en (3.20) avec $\boldsymbol{\Sigma_L}$ telle que définie dans l'équation (1.16). Les Tableaux 4.2 et 4.3 présentent les valeurs de l'estimation empirique de l'erreur de type I des tests CBM-RV et MURAT pour différentes valeurs du seuil α lorsque les phénotypes sont respectivement de lois marginales Normales et Gamma. Les diagrammes quantile-quantile pour la statis-



(a) (b)

Figure 4.2: Diagrammes quantile-quantile des p-valeurs du test CBM-RV pour 10 000 simulations sous H_0 sans correction pour l'estimation des paramètres et pour Y_1 et Y_2 de lois marginales : (a) Normales (b) Gamma.

tique du test CBM-RV sous H_0 sont présentés dans la Figure 4.3. Les diagrammes quantile-quantile pour les p-valeurs des scores CBM-RV et MURAT sous H_0 sont présentés dans les Figures 4.4 et 4.5 .

Tout d'abord, on constate qu'il y a inflation de l'erreur de type I pour le test de MURAT lorsque Y_1 et Y_2 suivent des lois Gamma, tel qu'illustré dans le diagramme quantile-quantile de la Figure 4.5. Par contre, l'erreur de type I est très bien contrôlée pour le test CBM-RV lorsqu'on applique la correction pour l'estimation des paramètres de nuisance. En effet, tel qu'illustré dans la Figure 4.3, peu importe les distributions marginales suivies par Y_1 et Y_2, la distribution observée pour la

Tableau 4.2: Estimation de l'erreur de type I des tests CBM-RV et MURAT avec correction pour l'estimation des paramètres de nuisance. Y_1 et Y_2 suivent des lois marginales Normales.

	$H_0 : \tau = 0$		
α	0.01	0.05	0.1
CBM-RV	0.010	0.052	0.105
MURAT	0.008	0.046	0.094

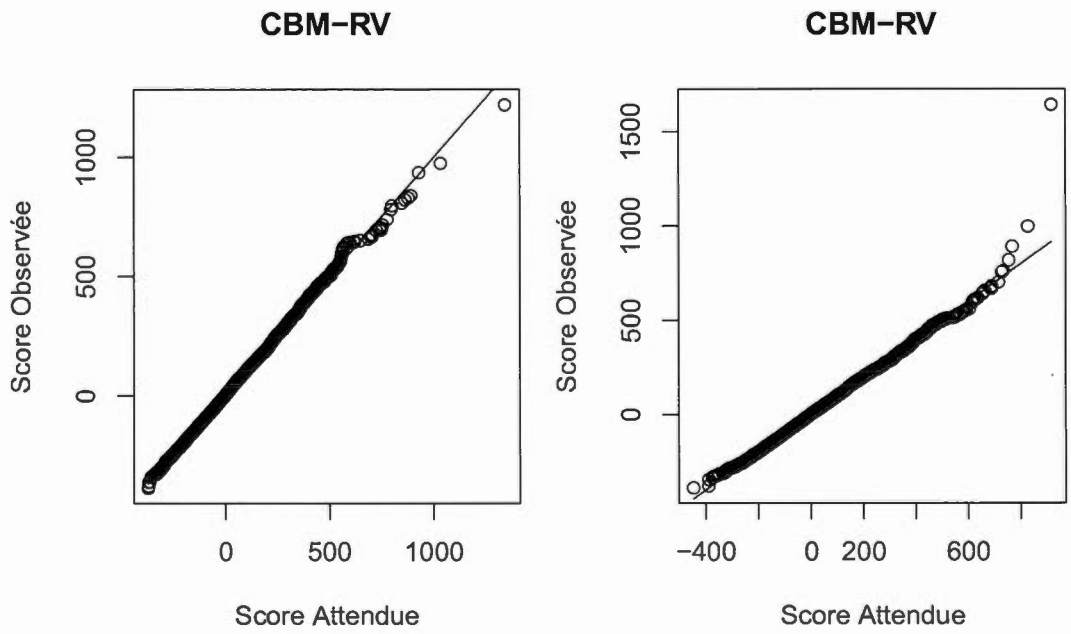
Tableau 4.3: Estimation de l'erreur de type I des tests CBM-RV et MURAT avec correction pour l'estimation des paramètres de nuisance. Y_1 et Y_2 suivent des lois marginales Gamma.

	$H_0 : \tau = 0$		
α	0.01	0.05	0.1
CBM-RV	0.0091	0.045	0.098
MURAT	0.032	0.087	0.141

statistique de score suit la distribution théorique attendue.

4.3 Scénario 2 : Erreur de spécification du modèle

On s'intéresse à l'inflation de l'erreur de type I du test CBM-RV lorsqu'on modélise la relation entre Y_1 et Y_2 par les copules de Clayton, de Frank et de Gumbel-Hougard, présentées dans le Tableau 2.1 du Chapitre 2. Le Tableau 4.4 présente les valeurs de l'estimation empirique de l'erreur de type I du test CBM-RV pour différentes valeurs du seuil α lorsqu'on spécifie une copule différente de la copule gaussienne. Les Figures 4.6 et 4.7 présentent les diagrammes quantile-quantile des p-valeurs du test CBM-RV sous H_0 pour chaque copule.



(a) (b)

Figure 4.3: Diagramme quantile-quantile de la statistique du test CBM-RV pour 10 000 simulations sous H_0 avec correction pour l'estimation des paramètres de nuisance et pour Y_1 et Y_2 de lois marginales : (a) Normales (b) Gamma.

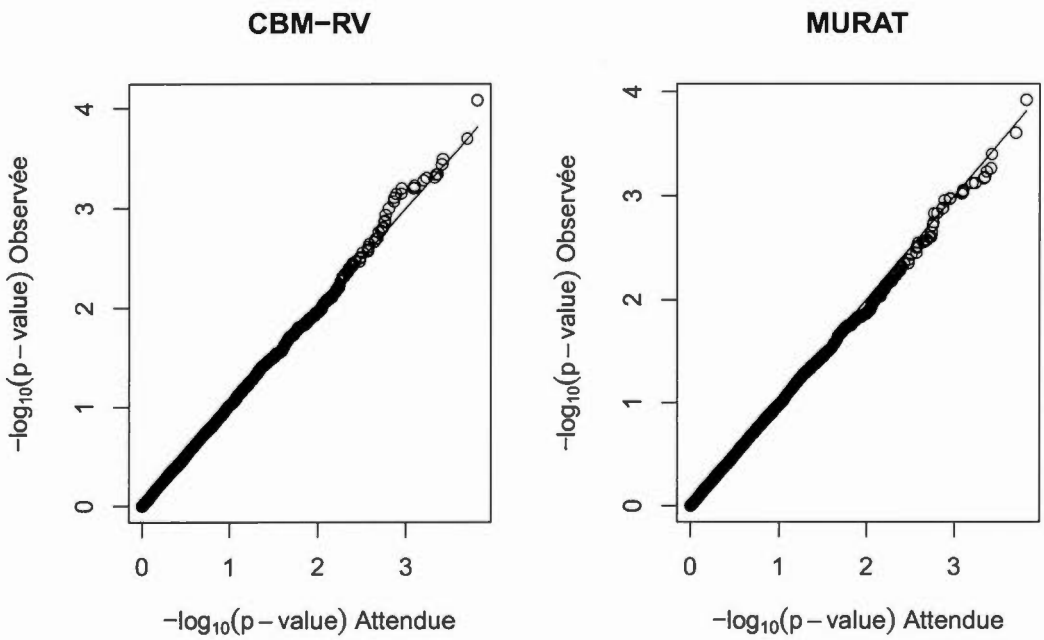


Figure 4.4: Diagrammes quantile-quantile des p-valeurs des tests CBM-RV et MURAT pour 10 000 simulations sous H_0 avec correction pour l'estimation des paramètres de nuisance et pour Y_1 et Y_2 de lois marginales Normales.

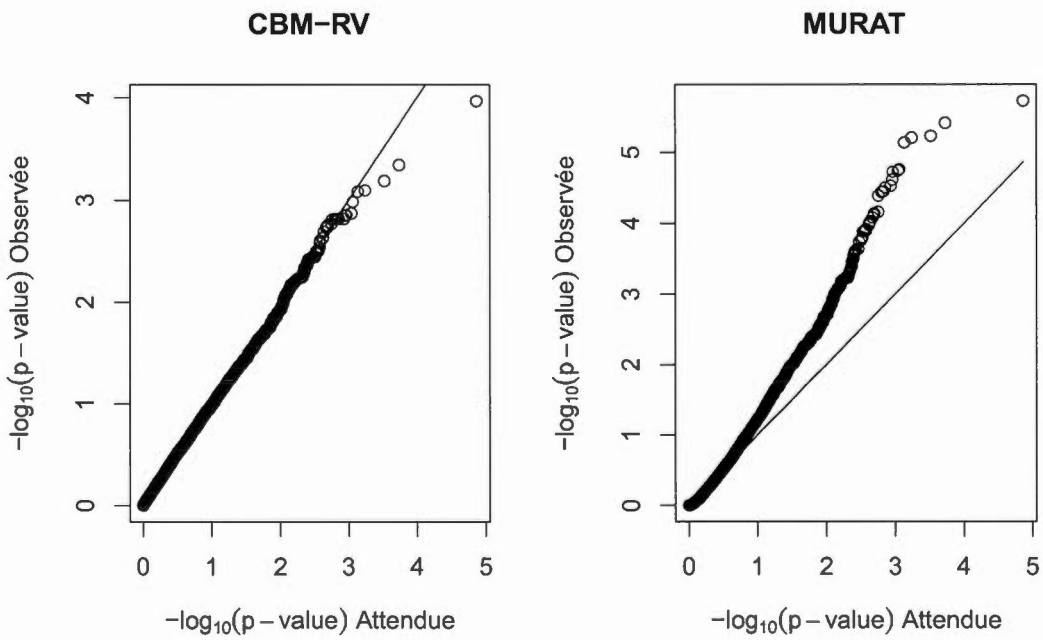


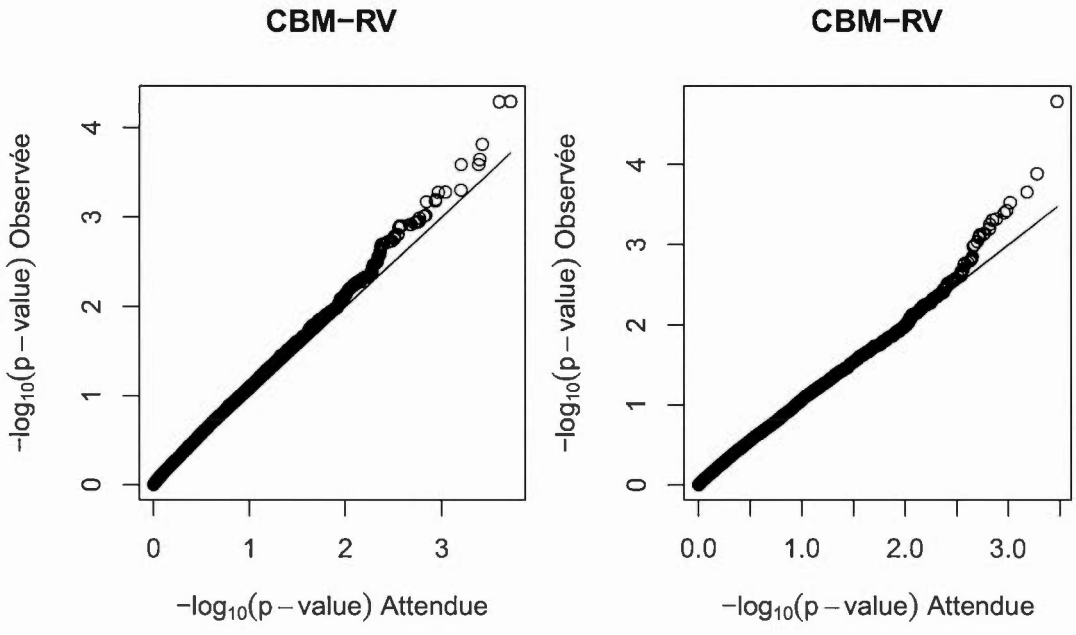
Figure 4.5: Diagrammes quantile-quantile des p-valeurs des tests CBM-RV et MURAT pour 10 000 simulations sous H_0 avec correction pour l'estimation des paramètres de nuisance et pour Y_1 et Y_2 de lois marginales Gamma.

On remarque que l'inflation de l'erreur de type I est moins importante lorsqu'on ajuste le modèle avec la copule de Frank qu'avec les copules de Clayton ou de Gumbel. Ceci est expliqué par le fait que la copule de Frank est symétrique, contrairement aux copules de Clayton et de Gumbel. En effet, tel que représenté dans la Figure C.1 en Annexe C, la dépendance est plus importante dans la queue négative pour la copule de Clayton, tandis que pour la copule de Gumbel la dépendance est plus importante dans la queue positive. La copule gaussienne étant symmétrique, on observe donc une inflation de l'erreur de type I plus importante lorsqu'on ajuste le modèle avec une copule asymétrique.

Enfin, on s'intéresse à l'inflation de l'erreur de type I du test CBM-RV lorsqu'il y a erreur de spécification sur les distributions marginales plutôt que sur le choix de la copule. Le Tableau 4.5 présente les valeurs pour l'estimation de l'erreur de type I lorsqu'on simule des traits selon des lois marginales Gamma et qu'on ajuste le modèle pour des lois Normales. La Figure 4.8 présente le diagramme quantile-quantile des p-valeurs du test CBM-RV pour ce scénario. Tel qu'attendu, l'inflation de l'erreur de type I est beaucoup plus importante lorsqu'il y a une erreur de spécification pour les distributions marginales des phénotypes que lorsque la distribution conjointe, c'est-à dire la copule, est mal spécifiée. Le choix de la copule pour modéliser la dépendance bivariée peut se baser sur le critère d'information d'Aikake (AIC), le meilleur modèle étant celui pour lequel la valeur d'AIC est la plus petite (Aikake, 1974).

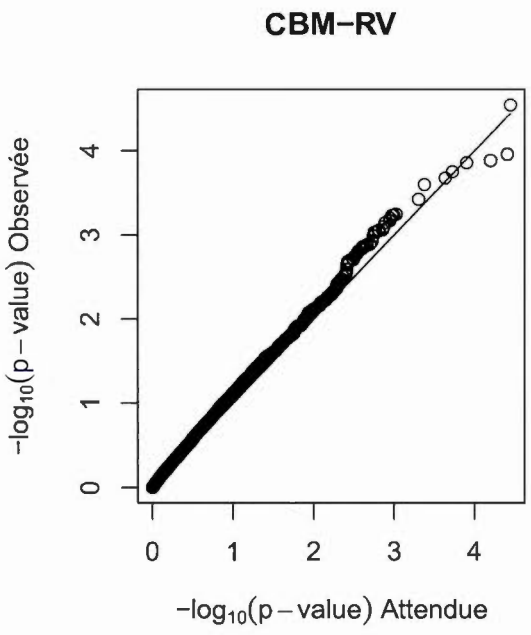
4.4 Scénario 3 : Puissance du test CBM-RV sous $H_A(\tau \neq 0)$

Pour évaluer la puissance du test CBM-RV sous l'hypothèse alternative, on s'intéresse aux situations pour lesquelles respectivement 10% et 20% des variants sont causaux. Soient $\{\nu_1, \nu_2, ..., \nu_u\}$ les u variants causaux choisis parmi les r variants rares, alors on pose $\beta_{1j} = \beta_{2j} = 0$ pour $j \notin \{\nu_1, \nu_2, ..., \nu_u\}$. Pour simuler les valeurs



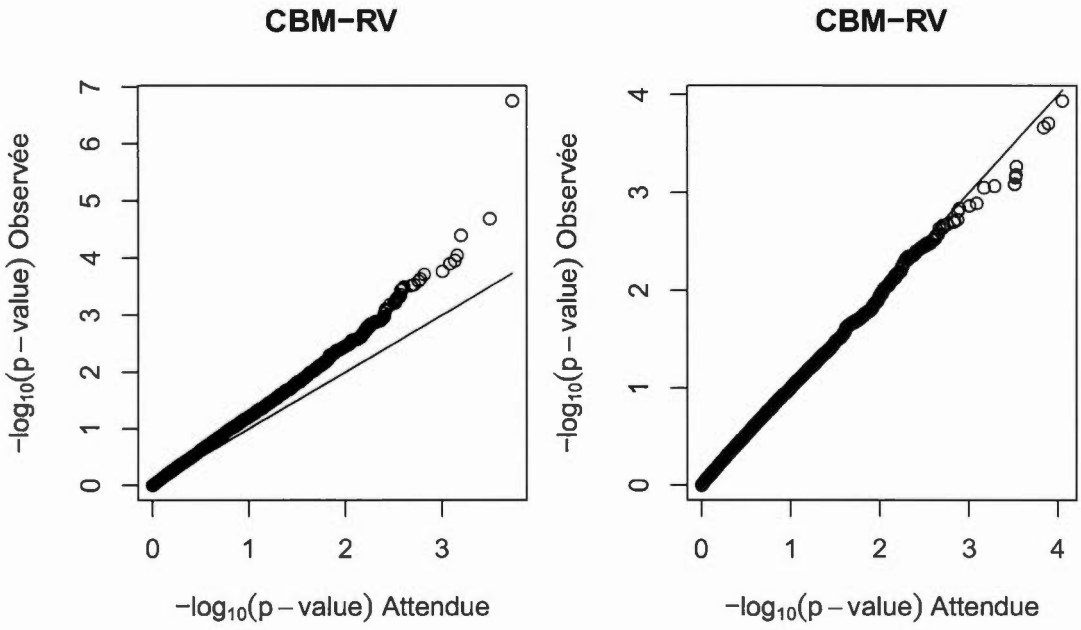
(a)

(b)



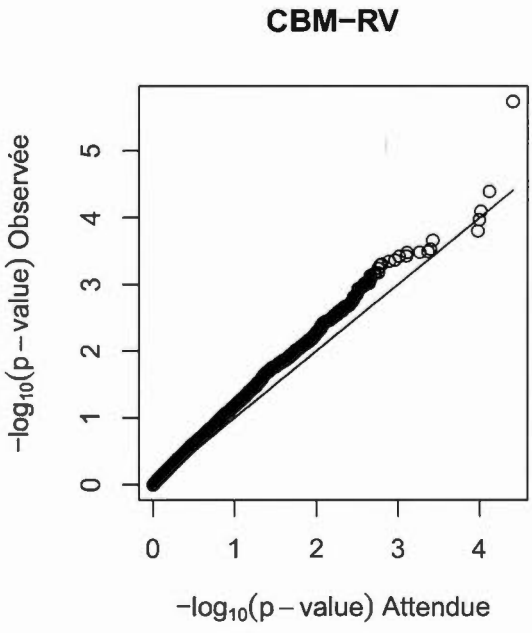
(c)

Figure 4.6: Diagrammes quantile-quantile des p-valeurs du test CBM-RV pour 10 000 simulations sous H_0 lorsque Y_1 et Y_2 sont de lois Normales et qu'on modélise la dépendance par la copule de : (a) Clayton (b) Frank (c) Gumbel-Hougard.



(a)

(b)



(c)

Figure 4.7: Diagrammes quantile-quantile des p-valeurs du test CBM-RV pour 10 000 simulations sous H_0 lorsque Y_1 et Y_2 sont de lois Gamma et qu'on modélise la dépendance par la copule de : (a) Clayton (b) Frank (c) Gumbel-Hougard.

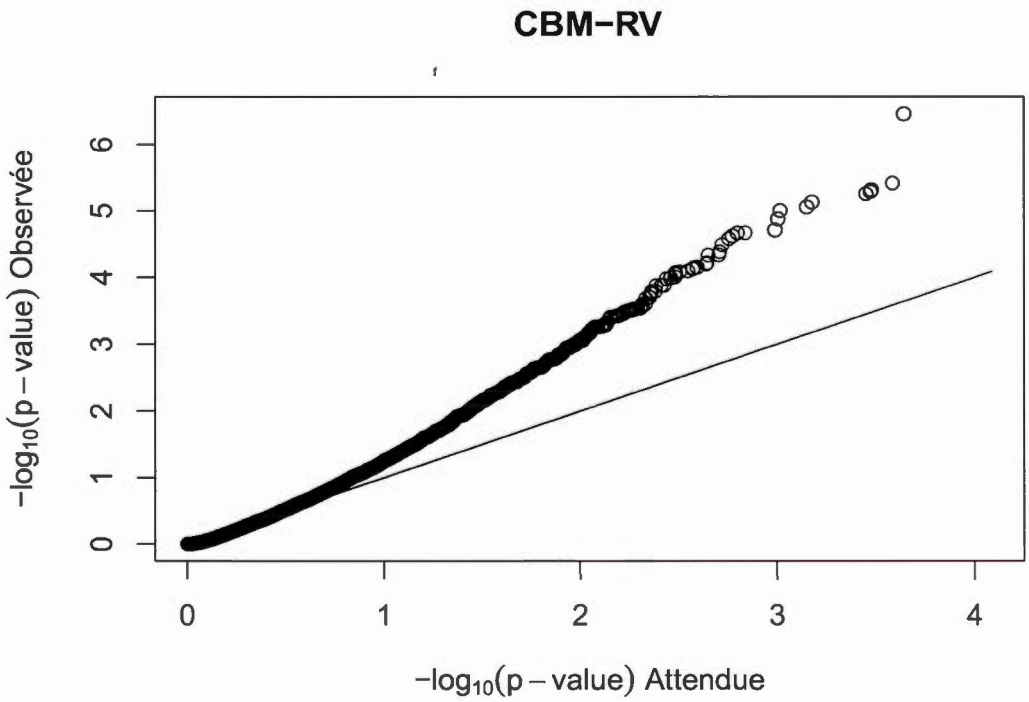


Figure 4.8: Diagrammes quantile-quantile des p-valeurs du test CBM-RV pour 10 000 simulations sous H_0 lorsque Y_1, Y_2 sont de lois Gamma et qu'on ajuste le modèle pour des lois normales.

Tableau 4.4: Estimation de l'erreur de type I du test CBM-RV lorsqu'il y a erreur de spécification pour la copule liant Y_1 et Y_2.

	Lois Normales			Lois Gammas		
α	0.01	0.05	0.1	0.01	0.05	0.1
Clayton	0.013	0.066	0.129	0.021	0.082	0.150
Frank	0.011	0.057	0.117	0.009	0.048	0.101
Gumbel	0.014	0.067	0.132	0.019	0.078	0.147

Tableau 4.5: Estimation de l'erreur de type I du test CBM-RV lorsqu'il y a erreur de spécification pour les distributions marginales de Y_1 et Y_2.

α	0.01	0.05	0.1
	0.037	0.097	0.154

des β_j pour les variants causaux, il faut fixer une valeur de τ de façon à obtenir une héritabilité autour de 2%, où l'héritabilité est définie comme la proportion de la variation phénotypique expliquée par la région génomique. Dans la prochaine section, on présente comment calculer τ en fonction de l'héritabilité du modèle.

4.4.1 Héritabilité

Typiquement, pour évaluer la puissance statistique avec des études de simulation, on cherche à avoir une héritabilité des traits autour de 2%. On définit l'héritabilité d'un trait comme la fraction de la variabilité phénotypique totale observée qui est attribuable à la variabilité génétique. Pour le modèle linéaire généralisé mixte de

l'équation (3.3), l'héritabilité pour le k^e trait est donnée par

$$\begin{aligned} h_k^2 &= \frac{\text{Var}[\boldsymbol{G}_i^T \boldsymbol{\beta}_k]}{\text{Var}[Y_{ik}]} \\ &= \frac{\text{Var}[\text{E}[\boldsymbol{G}_i^T \boldsymbol{\beta}_k | \boldsymbol{G}_i]] + \text{E}[\text{Var}[\boldsymbol{G}_i^T \boldsymbol{\beta}_k | \boldsymbol{G}_i]]}{\text{Var}[\text{E}[Y_{ik} | \boldsymbol{X_i}, \boldsymbol{G_i}, \boldsymbol{\beta}_k]] + \text{E}[\text{Var}[Y_{ik} | \boldsymbol{X_i}, \boldsymbol{G_i}, \boldsymbol{\beta}_k]]} \\ &= \frac{\text{E}[\tau \ \boldsymbol{G}_i^T \boldsymbol{W} \boldsymbol{G}_i]}{\text{Var}[g_k^{-1}(\boldsymbol{X}_i^T \boldsymbol{\gamma_j} + \boldsymbol{G_i}^T \boldsymbol{\beta}_k)] + \text{E}[\phi_k \nu(\mu_{ik})]} \end{aligned} \tag{4.1}$$

puisque $\boldsymbol{\beta_k} \sim F(\boldsymbol{0}_r, \tau \boldsymbol{W})$.

Il est difficile d'évaluer le dénominateur de l'équation (4.1) dans le cas où le modèle est non linéaire, comme par exemple pour un trait dont la distribution marginale est de loi Gamma, avec $g^{-1}(\cdot) = \exp(\cdot)$ et $\nu(\mu_i) = \mu_i^2$. Dans ce cas, il faut se résoudre à calculer l'héritabilité de manière empirique à l'aide de simulations. Dans le cas des modèles linéaires, l'équation (4.1) se simplifie

$$\begin{aligned} h_k^2 &= \frac{\tau \cdot \text{E}\left[\sum_{j=1}^r w_j g_{ij}^2\right]}{\text{Var}[\boldsymbol{X}_i^T \boldsymbol{\gamma_k} + \boldsymbol{G_i}^T \boldsymbol{\beta}_k] + \phi_k} \\ &= \frac{\tau \cdot \sum_{j=1}^r w_j \text{E}[g_{ij}^2]}{\tau \cdot \sum_{j=1}^r w_j \text{E}[g_{ij}^2] + \phi_k} \\ &= \frac{2\tau \cdot \sum_{j=1}^r w_j p_j (p_j + 1)}{2\tau \cdot \sum_{j=1}^r w_j p_j (p_j + 1) + \phi_k} \end{aligned} \tag{4.2}$$

puisque $g_{ij} \sim \text{Binomiale}(2, p_j)$ avec p_j la fréquence de l'allèle mineure pour le j^e locus, $j = 1, .., r$.

On isole le coefficient τ dans l'équation (4.2), d'où

$$\tau = \frac{\phi_k h_k^2}{2(1 - h_k^2) \sum_{j=1}^r w_j p_j (p_j + 1)}. \tag{4.3}$$

On pose $\phi_k = 1$, $h_k^2 = 0.02$ et

$$w_j = \begin{cases} 1, & j \in \{\nu_1, \nu_2, ..., \nu_u\} \\ 0, & \text{sinon} \end{cases}$$

dans l'équation (4.2), ce qui donne

$$\tau = \frac{1}{98\sum_{j\in\{\nu_1,\nu_2,\ldots,\nu_u\}} p_j(p_j+1)} \approx \frac{1}{98\sum_{j\in\{\nu_1,\nu_2,\ldots,\nu_u\}} p_j} \tag{4.4}$$

étant donné que $p_j < 0.05$ pour les variants rares. L'équation (4.4) est utile pour trouver la valeur de τ pour laquelle l'héritabilité pour les phénotypes simulés est d'approximativement 2%. On constate que le coefficient τ est inversement proportionnel aux nombres de variants causaux ainsi qu'à leurs fréquences alléliques.

4.4.2 Résultats

On teste l'impact sur la puissance du test lorsque la corrélation ρ entre les effets d'un même variant sur différents phénotypes est grande. En effet, puisqu'on pose $\rho = 0$ dans notre modèle afin de contourner les difficultés reliées à l'estimation de ce paramètre, il est d'intérêt d'évaluer la puissance du test lorsque $\rho \neq 0$ dans les données simulées. Ainsi, on évalue la puissance du test CBM-RV pour un seuil de test $\alpha = 0.05$ et pour deux traits de lois marginales Normales, dont la dépendance est simulée selon la copule gaussienne avec $\rho_e = 0, 0.3, 0.6$ et $\rho = 0, 0.8$. On compare avec la puissance obtenue pour les tests MURAT et SKAT, pour lequel les p-valeurs ajustées sont égales à deux fois le minimum des p-valeurs univariées. Les résultats sont présentés dans les diagrammes en bâtons des Figure 4.9 et 4.10, respectivement pour $\rho = 0$ et $\rho = 0.8$.

Pour une faible corrélation résiduelle ρ_e, la corrélation pléiotropique ρ a peu d'effet sur la puissance des tests CBM-RV et MURAT, qui sont comparables. Cependant, pour $\rho = 0.8$, la puissance des tests diminue lorsque ρ_e augmente. Cela n'est pas surprenant, car lorsque ρ_e et ρ sont tous deux grands, les phénotypes sont fortement corrélés. Par conséquent, dans des situations de très forte corrélation, un test multivarié peut gagner peu de puissance par rapport à l'analyse d'un seul

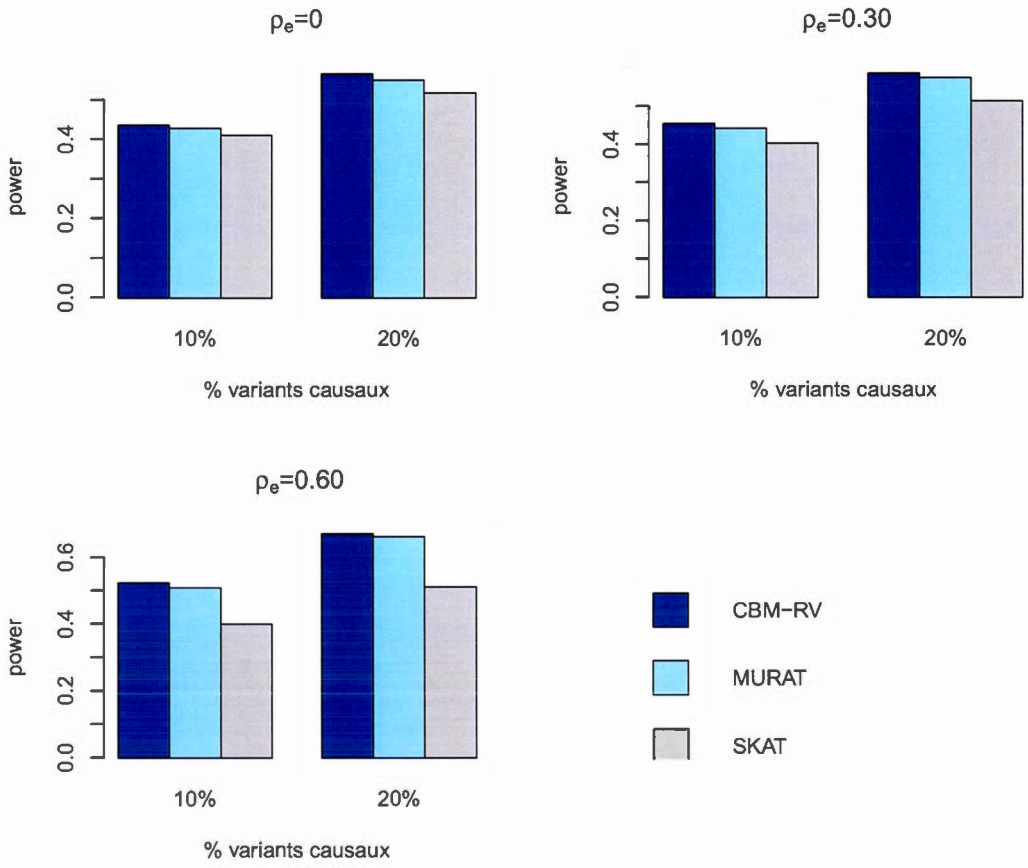


Figure 4.9: Puissance des tests CBM-RV, MURAT et SKAT lorsque $\rho = 0$

trait. Puisque MURAT estime la valeur de ρ tandis que notre modèle suppose $\rho = 0$, le test CBM-RV est moins puissant lorsque ρ et ρ_e sont grands. Enfin, pour toutes les valeurs de ρ_e, la puissance de SKAT diminue lorsque ρ augmente.

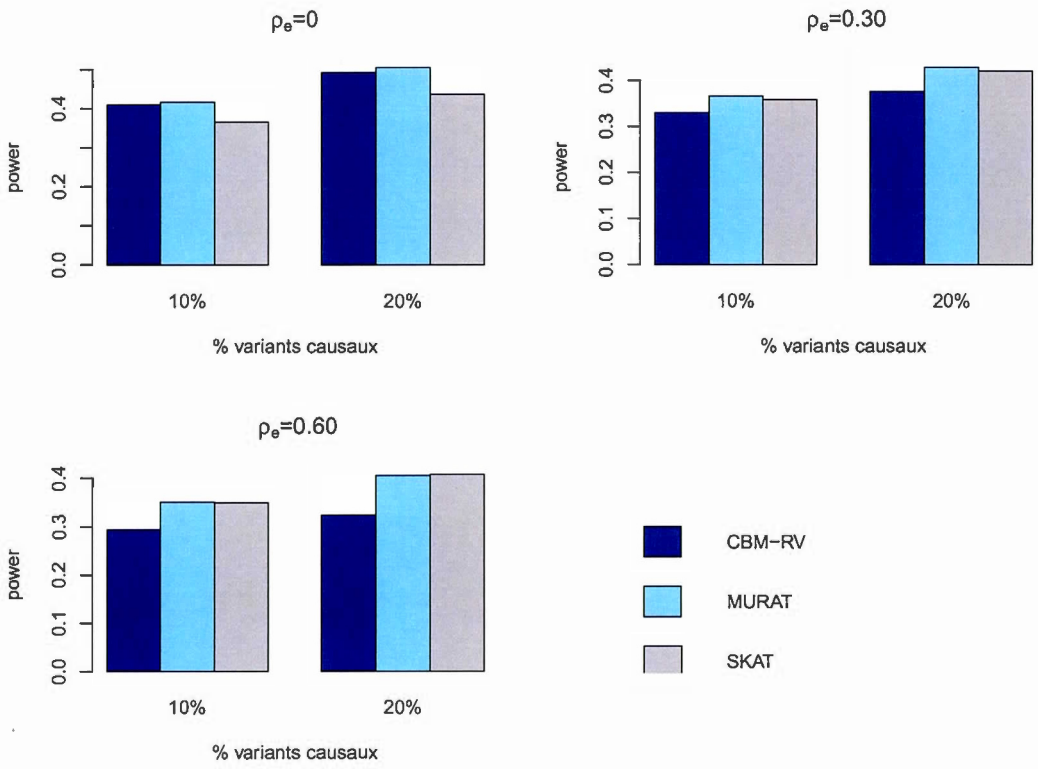


Figure 4.10: Puissance des tests CBM-RV, MURAT et SKAT lorsque $\rho = 0.8$

CHAPITRE V

ANALYSE DE DONNÉES RÉELLES

Dans ce chapitre, on illustre notre méthode sur les données de l'étude ALSPAC, pour *Avon Longitudinal Study of Parents and Children*. Cette étude prospective transgénérationelle s'intéresse aux facteurs génétiques, épigénétiques, biologiques, psychologiques et sociaux influençant divers indicateurs de santé et de développement social au cours d'une vie (Boyd *et al.*, 2013). On présente tout d'abord les données de l'étude, avant de comparer notre méthode avec les tests d'association SKAT et MURAT.

5.1 Présentation des données

La cohorte est composée d'enfants nés entre les années 1990 et 1992 dans la région de Bristol au Royaume-Uni. Pour l'analyse, on s'intéresse à 1477 sujets pour lesquels le génome entier a été séquencé. De plus, plusieurs phénotypes cliniques ont été mesurés pour ces sujets, dont les lipoprotéines de haute densité (HDL) et de basse densité (LDL), responsables du transport du cholestérol vers le foie, les triglycérides (TG), les apolipoprotéines B (ApoB) et les apolipoprotéines A1 (ApoA1) qui sont respectivement les constituants principaux des lipoprotéines de basse densité et de haute densité. On s'intéresse à l'association entre ces phénotypes et les gènes APOC3 et APOA1, situés sur le chromosome 11. Des mutations dans le gène APOA1 causent une diminution des niveaux de HDL dans le sang, ce

Tableau 5.1: Corrélation entre les phénotypes de la cohorte ALSPAC

Phénotypes		Corrélation
LDL	ApoB	0.873
HDL	ApoA1	0.828
HDL	Trigl	-0.408
ApoB	Trigl	0.188
HDL	ApoB	-0.183
ApoA1	Trigl	-0.127
ApoA1	ApoB	-0.073
LDL	Trigl	-0.036
HDL	LDL	-0.031
LDL	ApoA1	0.021

qui augmenterait potentiellement le risque de maladies cardiovasculaires (National Library of Medicine, 2012).

Le Tableau 5.1 présente les valeurs du coefficient de corrélation de Pearson pour chaque paire de phénotypes. Tel qu'attendu, la corrélation entre LDL et ApoB (0.873) et entre HDL et ApoA1 (0.828) est très élevée. De plus, HDL est aussi fortement corrélé avec Trigl. Ainsi, on considère les paires de phénotypes (HDL, ApoA1) et (HDL, Trigl) pour l'analyse d'association.

Il est important de vérifier la normalité bivariée pour les phénotypes (HDL, ApoA1) et (HDL, Trigl), puisqu'on a montré dans le chapitre précédent que le test de MURAT est sujet à l'inflation de l'erreur de type I si les marges ne sont pas normalement distribuées. Le Tableau 5.2 présente différentes statistiques de tests pour la normalité multivariée ainsi que les p-valeurs associées. Globalement, on rejette l'hypothèse de normalité bivariée pour les deux couples de phénotypes.

Tableau 5.2: Tests de normalité multivariée

	HDL & ApoA1		HDL & Trigl	
Test	Statistique	p-valeur	Statistique	p-valeur
Mardia				
Asymétrie	21.258	2.81×10^{-4}	65.082	2.47×10^{-13}
Aplatissement	-1.52	0.129	-5.72	1.08×10^{-8}
Henze-Zirkler	4.181	1.07×10^{-11}	10.16	0
Royston	72.173	1.30×10^{-16}	135.549	3.21×10^{-30}

Bien que la normalité bivariée pour les phénotypes (HDL, ApoA1) et (HDL, Trigl) est rejetée, ceci n'implique pas nécessairement que les distributions marginales des traits univariés soient non normales. D'après les p-valeurs issues des tests Shapiro-Wilk, incluses dans les histogrammes de la Figure 5.1, la normalité univariée est aussi rejetée pour chaque trait.

Afin d'éliminer l'inflation possible de l'erreur de type I des tests de MURAT et SKAT, on transforme les données de deux façons différentes. Tout d'abord, on applique une transformation logarithmique à chacun des trois traits en supposant qu'ils proviennent de lois log-normales. Les histogrammes et diagrammes quantile-quantile pour les phénotypes HDL, Trigl et ApoA1 après transformation logarithmique et ajustement pour le sexe sont présentés respectivement dans les Figure 5.2 et Figures D.1 à D.3 situées en Annexe D. Globalement, les p-valeurs pour le test de Shapiro-Wilk sont plus grandes qu'avant la transformation logarithmique, ce qui signifie que la distribution des résidus s'ajuste mieux à la loi log-normale. En regardant les diagrammes quantile-quantile de l'annexe D, on remarque que la normalité est tout de même rejetée à cause d'un nombre important d'observations dans les ailes de la distribution.

Une seconde conversion possible des données consiste à appliquer une tranformation normale inverse (TNI) sur les rédidus de chacun des traits, après ajustement pour le sexe (Beasley *et al.*, 2009). Soit r_i le rang pour le résidu de l'observation y_i, avec $i = 1, ..., n$, alors le score normal Y_i^t est obtenu par

$$Y_i^t = \Phi^{-1}\left(\frac{r_i}{n+1}\right), \tag{5.1}$$

avec Φ^{-1} la fonction quantile de la loi normale standard. En appliquant la fonction de répartition inverse de la loi normale standard sur la fonction de répartition empirique des observations, on s'assure que les résidus pour chaque trait sont maintenant distribués selon une loi $N(0, 1)$.

5.2 Tests d'association pour les gènes APOA1 et APOC3

Le premier génotype est formée de 27 SNPs situés sur le gène APOA1, dont 22 variants rares. Pour le gène APOC3, 43 variants rares pour un total de 59 SNPs composent le second génotype. Pour tester l'association entre les génotypes et les phénotypes (HDL, ApoA1) et (HDL, Trigl) avec notre méthode, on modélise les distributions marginales par des lois Gamma, ce qui est intéressant étant donnée la grande flexibilité de cette famille de distributions. De plus, on évite de cette façon de devoir transformer les variables avant l'analyse d'association. Pour chaque trait, on applique le test de Kolmogorv-Smirnov afin de tester l'hypothèse nulle que les données sont distribuées selon des lois Gamma. Le Tableau 5.3 présente les p-valeurs pour le test de Kolmogorov-Smirnov, après estimation des paramètres de la loi Gamma par maximum de vraisemblance. On peut considérer que la loi Gamma modélise convenablement la fonction de répartion aléatoire des traits. Le choix de la copule pour modéliser la dépendance bivariée est basée sur le critère d'information d'Aikake (AIC). On trouve que la copule gaussienne minimise l'AIC pour les deux paires de phénotypes. Enfin, on ajuste pour le sexe à l'aide d'une fonction de lien logarithmique.

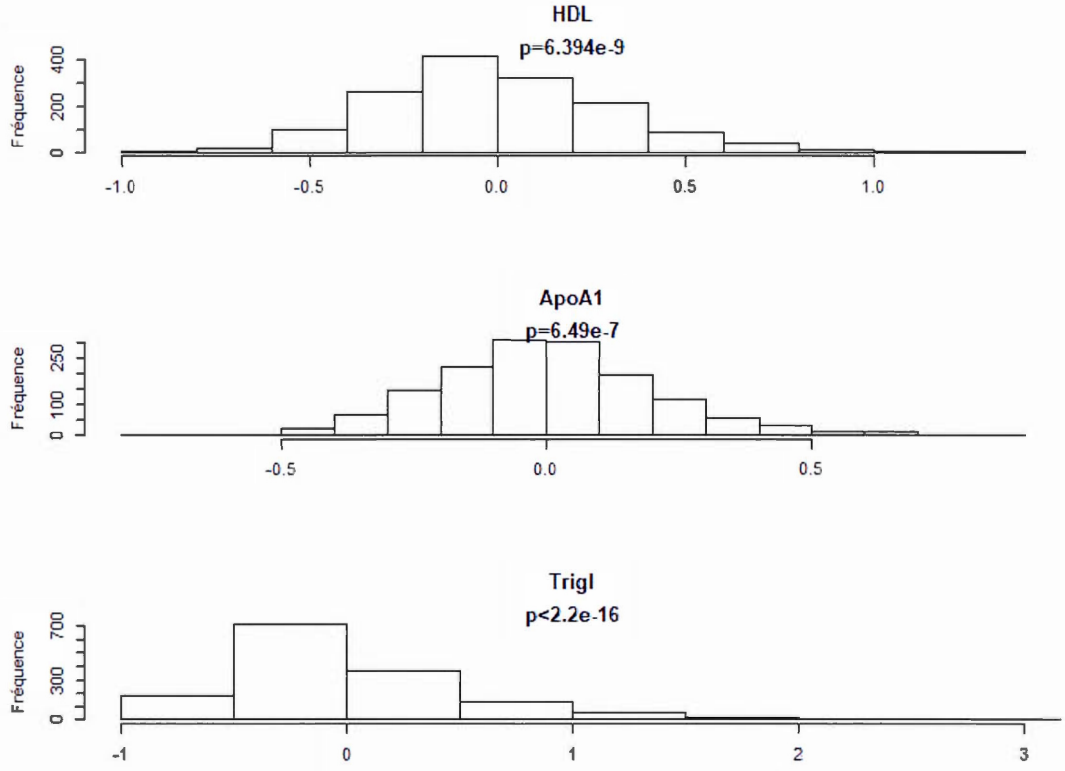


Figure 5.1: Histogrammes pour les phénotypes HDL, Trigl et ApoA1 après ajustement pour le sexe. Les p-valeurs proviennent du test de normalité Shapiro-Wilk.

Pour les tests SKAT et MURAT, on compare l'association avant et après transformation des données. Les p-valeurs pour les tests d'association SKAT, MURAT et CBM-RV sont présentées, respectivement pour les gènes APOA1 et APOC3, dans les Tableaux 5.4 et 5.5. Pour le test de SKAT, on applique la correction de Bonferroni, c'est-à-dire qu'on multiplie par deux la plus petite p-valeur entre les deux traits. La dernière colonne rapporte la valeur optimale de la corrélation pléiotropique, ρ_v, qui donne la p-valeur minimum pour le test de MURAT. On insère cette valeur dans notre modèle afin d'augmenter la puissance du test CBM-RV.

Dans le cas du phénotype bivarié (HDL,ApoA1), la valeur optimale pour la corré-

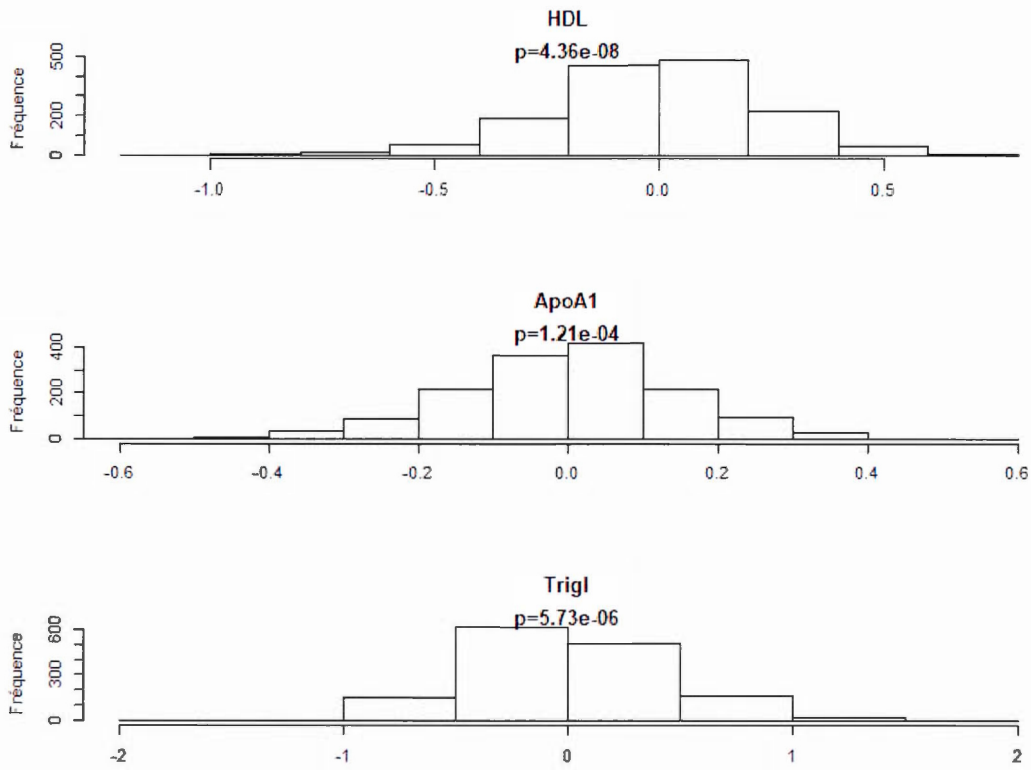


Figure 5.2: Histogrammes pour les phénotypes HDL, Trigl et ApoA1 après transformation logarithmique et ajustement pour le sexe. Les p-valeurs proviennent du test de normalité Shapiro-Wilk.

lation pléiotropique est estimée à $\rho_v = 0.8$. De plus, tel que mentionné précédemment, la corrélation entre HDL et ApoA1 est très élevée (0.828). Dans les études de simulation du chapitre précédent, on a démontré que lorsque les corrélations pléiotropiques et résiduelles entre deux traits sont élevées, les tests multivariés MURAT et CBM-RV ne gagnent pas en puissance comparativement au test univarié SKAT. Dans le cas du phénotype bivarié (HDL,Trigl), la corrélation est moins élevée entre les deux traits (-0.408), tandis que la corrélation pléiotropique qui optimise la puissance du test MURAT est estimée à $\rho_v = 0$. Ainsi, les tests

Tableau 5.3: Test de Kolmogorov-Smirnov

Trait	Distribution	p-valeur
HDL	Gamma($\alpha = 22.6, \beta = 0.06$)	0.39
ApoA1	Gamma($\alpha = 53, \beta = 0.03$)	0.56
Trigl	Gamma($\alpha = 7.09, \beta = 0.15$)	0.03

Tableau 5.4: P-valeurs des tests d'association génétique SKAT, MURAT et CBM-RV pour le gène APOA1

Traits	SKAT	MURAT	CBM-RV	ρ_v
(HDL,ApoA1)				
- Y_i	4.90×10^{-4}	8.07×10^{-3}	3.64×10^{-3}	0.8
- log Y_i	6.75×10^{-4}	9.48×10^{-3}	-	
- Y_i^t	0.580	0.519	-	
(HDL,Trigl)				
- Y_i	4.90×10^{-4}	1.48×10^{-6}	1.20×10^{-5}	0
- log Y_i	9.00×10^{-4}	7.42×10^{-7}	-	
- Y_i^t	0.580	0.365	-	

multivariés MURAT et CBM-RV gagnent en puissance comparativement au test SKAT.

La différence de puissance entre les tests MURAT et CBM-RV pourrait être causée par une inflation de l'erreur de type I pour le test de MURAT dû au fait que la normalité marginale des traits n'est pas respectée. D'ailleurs, on remarque que la puissance des tests SKAT et MURAT est grandement affectée lorsqu'on transforme les observations par la méthode de la TNI, comparativement à la transformation logarithmique. On pourrait expliquer cela par le fait qu'on perd l'infor-

mation contenue dans les valeurs extrêmes de la distribution lorsqu'on transforme les traits en scores normaux. L'avantage de notre test CBM-RV est le fait d'assumer des fonctions de répartitions marginales Gamma, ce qui évite les risques d'inflation de l'erreur de type I ainsi que les complications en termes d'analyse et d'interprétation inhérentes au fait de devoir transformer les variables.

Tableau 5.5: P-valeurs des tests d'association génétique SKAT, MURAT et CBM-RV pour le gène APOC3

Traits	SKAT	MURAT	CBM-RV	ρ_v
(HDL,ApoA1)				
- Y_i	3.32×10^{-4}	4.75×10^{-5}	2.83×10^{-3}	0.8
- log Y_i	2.86×10^{-4}	5.91×10^{-5}	-	
- Y_i^t	0.880	0.922	-	
(HDL,Trigl)				
- Y_i	3.57×10^{-4}	1.45×10^{-6}	1.35×10^{-5}	0
- log Y_i	6.24×10^{-4}	1.88×10^{-6}	-	
- Y_i^t	0.880	0.459	-	

CONCLUSION

Dans ce projet, notre objectif était de développer un test d'association génétique multivarié pour des phénotypes continus distribués selon des lois autres que la loi normale. Pour cela, nous avons intégré des copules à un modèle à effets aléatoires, afin de modéliser la structure de dépendance entre les différents traits indépendamment de leurs distributions marginales. Dans le but d'incorporer les variants rares au modèle, nous avons proposé un test statistique basé sur un test de score sur une composante de variance. Afin de trouver une forme analytique pour notre test, nous avons effectué une approximation de Taylor au deuxième degré de la vraisemblance du modèle sous l'hypothèse nulle.

À l'aide d'échantillons simulés, nous avons analysé le comportement de notre statistique de test sous différentes hypothèses. On a démontré que peu importe les distributions marginales des traits, l'erreur de type I est toujours contrôlée. De plus, nous avons observé que la puissance de notre test était comparable au test de MURAT pour les échantillons simulés. Enfin, nous avons été en mesure d'appliquer notre méthode sur des données réelles provenant de la cohorte ALSPAC.

Les résultats obtenus à partir de ce projet sont encourageants et suggèrent que le modèle proposé soit approfondi sur plusieurs facettes. Il serait intéressant, dans un premier lieu, d'évaluer la performance de notre modèle sur des phénotypes multivariés de dimension plus grande que deux. Ensuite, il serait souhaitable de pouvoir étendre notre méthode aux phénotypes discrets, ce qui n'est pas trivial, étant données les complications dues à l'usage des copules en présence de variables discrètes. Enfin, il serait pertinent d'appliquer notre test sur des données fami-

liales, pour lesquelles le phénotype multivarié serait formé à partir des traits de tous les individus provenant de la même famille.

ANNEXE A

DÉRIVATION PAR RAPPORT À UN VECTEUR

Soit $\boldsymbol{x}$ un vecteur de $\mathbb{R}^n$, $f(\boldsymbol{x})$ une fonction scalaire de $\boldsymbol{x}$ et $\boldsymbol{g}(\boldsymbol{x})$ une fonction vectorielle de x dans $\mathbb{R}^m$, alors on définit les propriétés suivantes :

a. La dérivée première de $f(\boldsymbol{x})$ par rapport à x est définie comme

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

b. La dérivée seconde de $f(\boldsymbol{x})$ par rapport à x est définie comme

$$\frac{\partial}{\partial \boldsymbol{x}\boldsymbol{x}^T} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

c. La dérivée première de $\boldsymbol{g}(\boldsymbol{x})$ par rapport à x est définie comme

$$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix}$$

ANNEXE B

DÉRIVÉES PREMIÈRE ET SECONDE DE LA LOG-VRAISEMBLANCE CONDITIONNELLE

La log-vraisemblance conditionnelle du modèle, telle qu'obtenue en (3.15), est donnée par

$$l(\boldsymbol{\theta}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\log f_1(y_{i1}|\boldsymbol{\beta}) + \log f_2(y_{i2}|\boldsymbol{\beta}) + \log c_\alpha(F_1(y_{i1}|\boldsymbol{\beta}); F_2(y_{i2}|\boldsymbol{\beta}))\right].$$

Puisque le modèle linéaire généralisé mixte posé en (3.3) suppose des densités marginales provenant de familles de lois exponentielles sous la forme de l'équation (1.8), on a

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_{ij}|\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} - c(y_i, \phi)\right] \\
&= \frac{1}{\phi}\left[y_{ij}\frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} - \frac{\partial b(\theta_{ij})}{\partial \theta_{ij}}\frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}}\right] \\
&= \frac{1}{\phi}[y_{ij} - \mu_{ij}]\frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}}
\end{aligned}$$

en utilisant le résultat de l'équation (1.9), soit $b'(\theta) = \mu$. On simplifie en utilisant les propriétés des dérivées en chaîne, ce qui donne

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_{ij}|\boldsymbol{\beta}) &= \frac{1}{\phi}[y_{ij} - \mu_{ij}] \frac{\partial \theta_{ij}}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{\phi}[y_{ij} - \mu_{ij}] \frac{\partial \mu_{ij}}{\partial \theta_{ij}}^{-1} \frac{\partial \mu_{ij}}{\partial g(\mu_{ij})} \frac{\partial g(\mu_{ij})}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{\phi}[y_{ij} - \mu_{ij}] \frac{\partial^2 b(\theta_{ij})}{\partial \theta_{ij}^2}^{-1} \frac{\partial g(\mu_{ij})}{\partial \mu_{ij}}^{-1} \frac{\partial g(\mu_{ij})}{\partial \boldsymbol{\beta}}\end{aligned}$$

en utilisant encore une fois le résultat de l'équation (1.9). On a défini la fonction variance en (1.10) telle que $v(\mu) = b''(\theta)$, d'où

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_{ij}|\boldsymbol{\beta}) &= \frac{1}{\phi}[y_{ij} - \mu_{ij}] v(\mu_{ij})^{-1} \frac{\partial g(\mu_{ij})}{\partial \mu_{ij}}^{-1} \begin{bmatrix} \frac{\partial g(\mu_{ij})}{\partial \beta_1} \\ \\ \frac{\partial g(\mu_{ij})}{\partial \beta_2} \end{bmatrix} \\ &= \frac{1}{\phi} \frac{[y_{ij} - \mu_{ij}]}{v(\mu_{ij}) \frac{\partial g(\mu_{ij})}{\partial \mu_{ij}}} \begin{bmatrix} \mathbb{1}_1(j) \\ \\ \mathbb{1}_2(j) \end{bmatrix} \otimes \boldsymbol{G}_i\end{aligned}$$

avec

$$\mathbb{1}_1(j) = \begin{cases} 1 & , \text{ si } j = 1 \\ \\ 0 & \text{autrement} \end{cases}.$$

Ainsi, le score en $\boldsymbol{\beta}$ pour la log-vraisemblance conditionnelle du modèle est

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}\begin{bmatrix} \frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} \\ \\ \frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i + \sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}}\log\, c_\alpha$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\begin{bmatrix} \frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} \\ \\ \frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i + \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}_1}\log\, c_\alpha \\ \\ \frac{\partial}{\partial \boldsymbol{\beta}_2}\log\, c_\alpha \end{bmatrix}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\begin{bmatrix} \frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} \\ \\ \frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i + \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}\frac{\partial \mu_{i1}}{\partial \boldsymbol{\beta}_1} \\ \\ \frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}\frac{\partial \mu_{i2}}{\partial \boldsymbol{\beta}_2} \end{bmatrix}$$

$$= \sum_{i=1}^{n}\begin{bmatrix} \frac{1}{\phi}\frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{\frac{\partial g(\mu_{i1})}{\partial \mu_{i1}}} \\ \\ \frac{1}{\phi}\frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{\frac{\partial g(\mu_{i2})}{\partial \mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i.$$

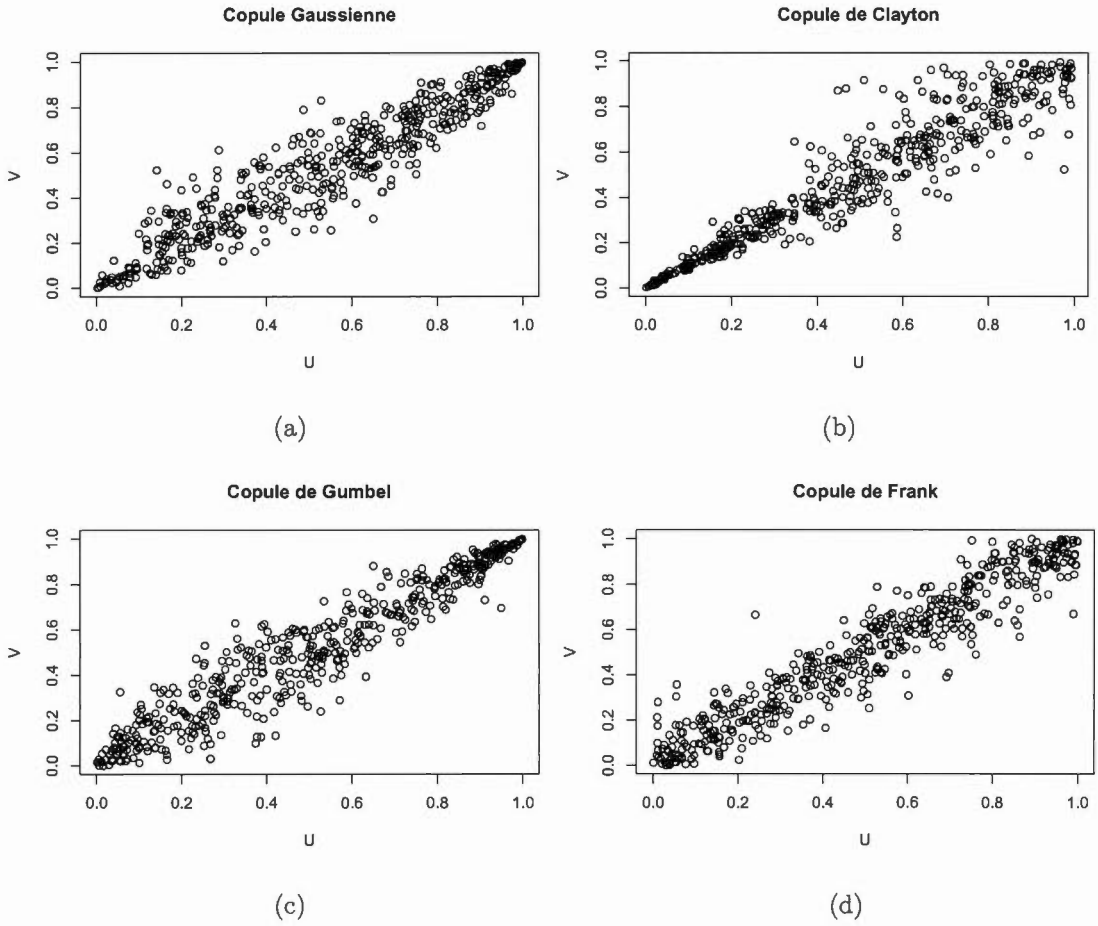
La dérivée double de la log-vraisemblance conditionnelle par rapport à $\boldsymbol{\beta}$ est donnée par

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} l(\boldsymbol{\theta}|\boldsymbol{\beta}) = \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}_1}\left[\frac{1}{\phi}\frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})g'_{\mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}\right] & \frac{\partial}{\partial \boldsymbol{\beta}_1}\left[\frac{1}{\phi}\frac{y_{i2}-\mu_{i2}}{v(y_{i2})g'_{\mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}\right] \\ \frac{\partial}{\partial \boldsymbol{\beta}_2}\left[\frac{1}{\phi}\frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})g'_{\mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}\right] & \frac{\partial}{\partial \boldsymbol{\beta}_2}\left[\frac{1}{\phi}\frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})g'_{\mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}\right] \end{bmatrix}^T \otimes \boldsymbol{G}_i^T$$

$$= \sum_{i=1}^{n}\begin{bmatrix} \frac{\partial}{\partial \mu_{i1}}\left[\frac{1}{\phi}\frac{y_{i1}-\mu_{i1}}{v(\mu_{i1})g'_{\mu_{i1}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}\right]\frac{1}{g'_{\mu_{i1}}} & \frac{\partial}{\partial \mu_{i1}}\left[\frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}\right]\frac{1}{g'_{\mu_{i1}}} \\ \frac{\partial}{\partial \mu_{i2}}\left[\frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i1}}}{g'_{\mu_{i1}}}\right]\frac{1}{g'_{\mu_{i2}}} & \frac{\partial}{\partial \mu_{i2}}\left[\frac{1}{\phi}\frac{y_{i2}-\mu_{i2}}{v(\mu_{i2})g'_{\mu_{i2}}} + \frac{\frac{\partial \log(c_\alpha)}{\partial \mu_{i2}}}{g'_{\mu_{i2}}}\right]\frac{1}{g'_{\mu_{i2}}} \end{bmatrix} \otimes \boldsymbol{G}_i\boldsymbol{G}_i^T.$$

ANNEXE C

DISTRIBUTION DE DIFFÉRENTES DE COPULES



(a) (b)

(c) (d)

Figure C.1: Distribution des différentes copules pour un τ de Kendall de 0.8

ANNEXE D

DIAGRAMMES QUANTILE-QUANTILE

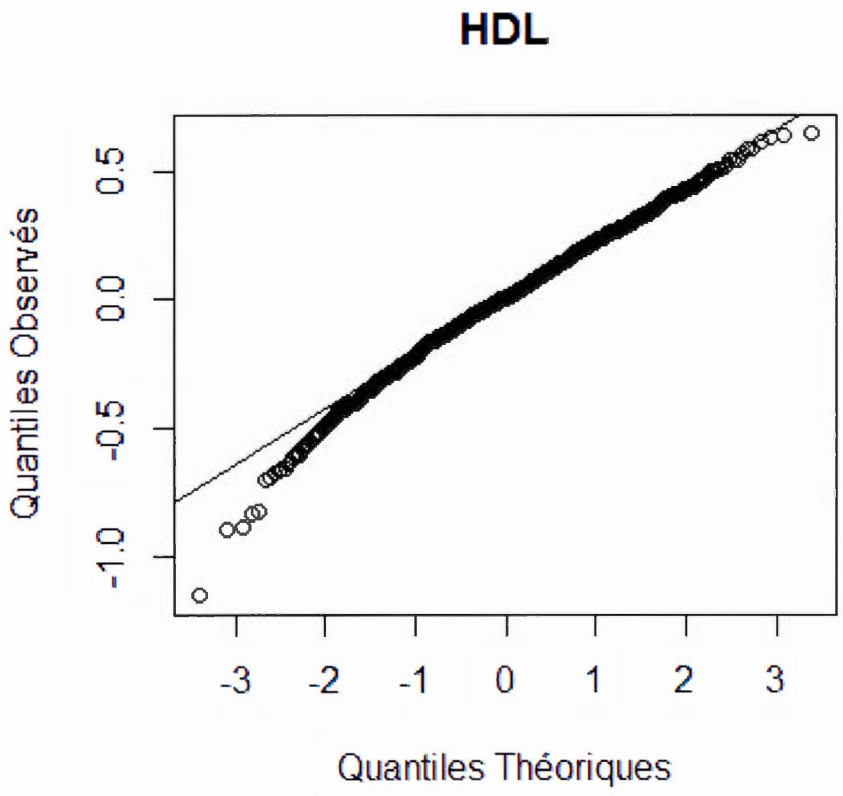


Figure D.1: Diagramme quantile-quantile pour le trait HDL après transformation logarithmique et ajustement pour le sexe.

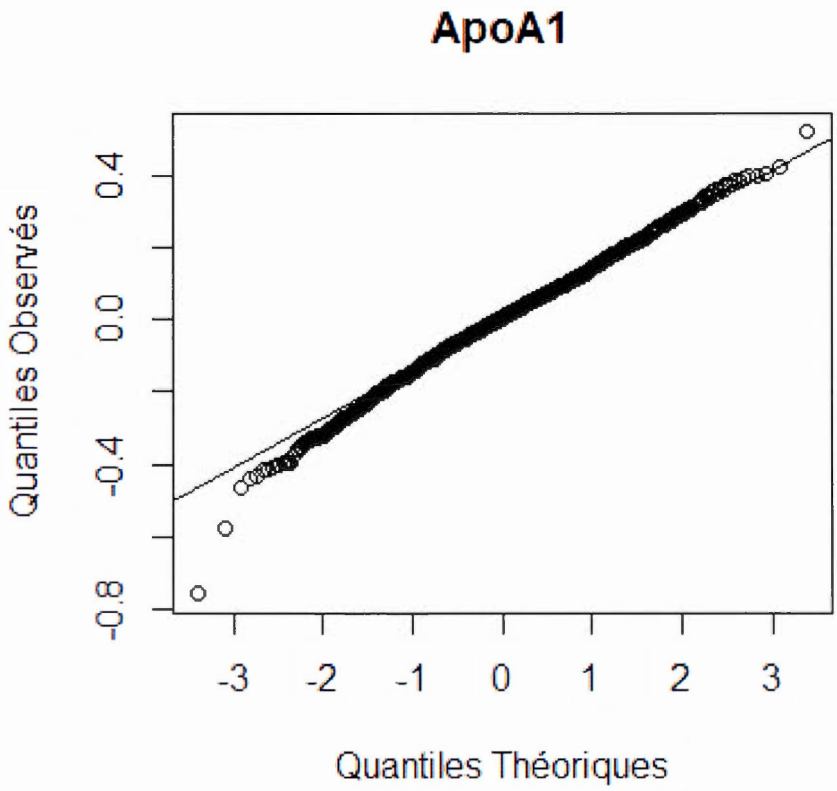


Figure D.2: Diagramme quantile-quantile pour le trait ApoA1 après transformation logarithmique et ajustement pour le sexe.

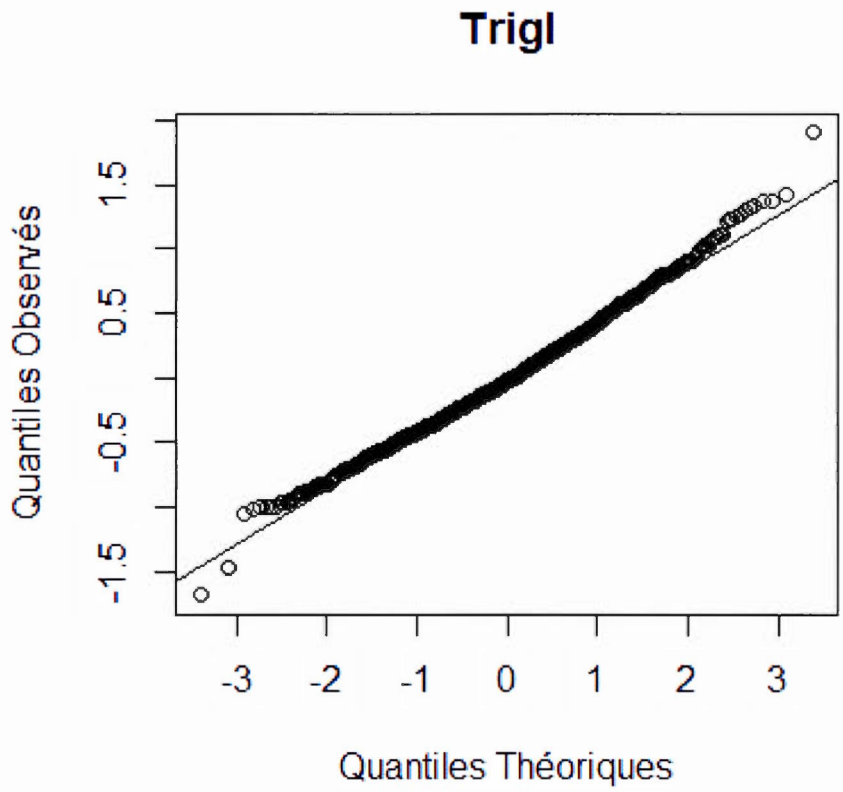


Figure D.3: Diagramme quantile-quantile pour le trait Trigl après transformation logarithmique et ajustement pour le sexe.

RÉFÉRENCES

Aikake, H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control.

Beasley, T. M., Erickson, S. et Allison, D. B. (2009). *Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited ?* Behavior Genetics.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A. et Henderson, J. (2013). *Cohort Profile : The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children.* International Journal of Epidemiology.

Fang, H.-B. et Fang, K.-T. (2002). *The Meta-elliptical Distributions with Given Marginals.* Journal of Multivariate Analysis.

IGSR (2017). The 1000 genomes project. [En ligne ; accédé le 20 Juillet 2017]. Récupéré de `http://www.internationalgenome.org/home`

Lee, S., Wu, M. C. et Lin, X. (2012). *Optimal tests for rare variant effects in sequencing association studies.* Biostatistics (Oxford, England).

Lin, X. (1997). *Variance component testing in generalised linear models with random effects.* Biometrika.

Lu, T.-T. et Shiou, S.-H. (2000). *Inverses of 2x2 Block Matrices.* Computers & Mathematics with Applications.

National Library of Medicine (2012). APOA1 gene. [En ligne ; accédé le 1er Août 2017]. Récupéré de `https://ghr.nlm.nih.gov/gene/APOA1`

National Library of Medicine (2015). BRCA1 gene. [En ligne ; accédé le 3 Octobre 2017]. Récupéré de `https://ghr.nlm.nih.gov/gene/BRCA1`

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer.

Rao, C. R. (2001). Two score and 10 years of score tests. *Journal of Statistical Planning and Inference*, *97*(1), 3 – 7. Rao's Score Test, `http://dx.doi.org/`

https://doi.org/10.1016/S0378-3758(00)00342-6. Récupéré de http://www.sciencedirect.com/science/article/pii/S0378375800003426

Sun, J., Oualkacha, K., Forgetta, V., Zheng, H.-F., Richards, J. B., Campia, A. et Greenwood, C. M. (2016). *A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects.* European Journal of Human Genetics.

Wang, X. et Yan, J. (2013). *Practical Notes On Multivariate Modeling Based on Elliptical Copulas.* Journal de la Société Française de Statistique.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. et Lin, X. (2011). *Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.* American Journal of Human Genetics.