Méthodes de factorisation matricielle pour la génomique des populations et les tests d'association

Kévin Caye

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Kévin CAYE

Thèse dirigée par **Olivier FRANÇOIS**
et co-encadrée par **Olivier MICHEL** et **Jean-Luc BOSSON**

préparée au sein du laboratoire **TIMC-IMAG**
et de l'école doctorale **"Ingénierie de la Santé, de la Cognition et Environnement" (EDISCE)**

Méthodes de factorisation matricielle pour la génomique des populations et les tests d'association

Thèse soutenue publiquement le 11 décembre 2017,
devant le jury composé de :
Michael BLUM
DR CNRS, UGA Grenoble, Président
Christophe AMBROISE
Professeur, UEVE Évry, Rapporteur
Charles BOUVEYRON
Professeur, UCA Nice, Rapporteur
Thomas BURGER
CR CNRS, CEA Grenoble, Examinateur

Table des matières

Remerciements

Tout d'abord je tiens à remercier Olivier François, mon directeur de thèse, à la fois pour son enseignement infini et pour sa patience (infinie aussi) quand il a du relire ce manuscrit. Je remercie toute l'équipe BCM pour la bonne ambiance de ces trois dernières années. Une mention spéciale pour Michael Blum, si nous n'avons jamais manqué de bière lors d'une session poster c'est bien grâce à lui.

Je remercie mon encadrant de thèse Olivier Michel, avec qui nos échanges ont toujours été très bénéfiques. Je remercie également Jean Luc Bosson, mon troisième, encadrant de thèse. Je remercie enfin mes rapporteurs Christophe Ambroise et Charles Bouveyron, ainsi que les membres du jury Thomas Burger et Michel Blum, qui ont accepté d'évaluer mon travail.

Je remercie l'ensemble de ma famille qui m'a toujours soutenu. En particulier mes deux petits frères et ma mère. Elle m'aura aidé à faire mes devoirs de la maternelle à la thèse.

Pour finir, merci ma puce, rien n'aurait été possible sans toi.

Résumé court

Titre : Méthodes de factorisation matricielle pour la génomique des populations et les tests d'association.

Résumé : Nous présentons des méthodes statistiques reposant sur des problèmes de factorisation matricielle. Une première méthode permet l'inférence rapide de la structure de populations à partir de données génétiques en incluant l'information de proximité géographique. Une deuxième méthode permet de corriger les études d'association pour les facteurs de confusion. Nous présentons dans ce manuscrit les modèles statistiques, ainsi que des aspects théoriques des algorithmes d'inférence. De plus, à l'aide de simulations numériques, nous comparons les performances de nos méthodes à celles des méthodes existantes. Enfin, nous appliquons nos méthodes sur des données biologiques réelles. Nos méthodes ont été implémentées et distribuées sous la forme de packages R : `tess3r` et `lfmm`.

Mots-clés : optimisation, factorisation de matrices, structure génétique des populations, métissage, étude d'association, facteurs de confusion, bio-informatique.

Title : Matrix factorization methods for population genomics and association mapping.

Summary : We present statistical methods based on matrix factorization problems. A first method allows efficient inference of population structure from genetic data and including geographic proximity information. A second method corrects the association studies for confounding factors. We present in this manuscript the models, as well as the theoretical aspects of the inference algorithms. Moreover, using numerical simulations, we compare the performance of our methods with those of existing methods. Finally, we use our methods on real biological data. Our methods have been implemented and distributed as R packages : `tess3r` and `lfmm`.

Keyword : optimization, matrix factorization, genetic structure of populations, admixture, association study, confounding factors, bioinformatics.

Chapitre 1

Introduction

Cette dernière décennie a été marquée par une accumulation des données dans tous les domaines de la science. Cette accumulation de données est une aubaine pour les scientifiques. En génétique, l'abondance des données permet de comprendre toujours mieux le vivant. Cependant, que faire d'autant de données et comment en tirer l'information qui permettra de mieux comprendre le monde qui nous entoure ? Il s'agit là d'un défi majeur pour les statistiques.

Le premier problème qui se pose avec les grands jeux de données est celui de la vitesse des algorithmes utilisés pour les traiter. Si nous avons beaucoup de données rapidement nous voulons aussi des algorithmes rapides pour les analyser et les visualiser. L'augmentation de la puissance des ordinateurs ne suffit pas toujours à rendre applicable un algorithme à un plus grand jeu de données. Par conséquent, il est toujours nécessaire de trouver des nouveaux algorithmes afin de traiter rapidement des données plus grandes.

Un deuxième problème important avec les grands échantillons de données concerne les études de liaisons entre variables. Il est bien connu des statisticiens qu'un lien statistique entre variables ne correspond pas nécessairement à un lien de causalité. Nous pouvons par exemple citer l'effet de Yule-Simpsons, un paradoxe statistique dans lequel un phénomène observé de plusieurs groupes semble s'inverser lorsque les groupes sont combinés (Simpson, 1951). Historiquement, les études de corrélation étaient faites sur des échantillons de données construits pour l'occasion et contrôlés pour les facteurs de confusion. Cependant, dans le contexte actuel les scientifiques ont accès à beaucoup de données provenant de sources éparses. Utiliser de telles données afin d'identifier des liens statistiques pertinents pour une problématique scientifique,

est alors beaucoup plus complexe qu'avec des données contrôlées. À cela s'ajoute le fait que plus l'échantillon de données est grand, plus il est possible de détecter des liens statistiques subtils. Par conséquent, dans le contexte actuel où l'accès aux données est très facile, il faut être très prudent et rigoureux dans les études de corrélation.

Dans le cadre de cette thèse de doctorat, nous nous sommes intéressés à développer des méthodes statistiques adaptées au traitement de données génétiques afin de répondre à deux problèmes : l'estimation de la structure génétique des populations à partir de données génétiques et géographiques, et l'estimation des facteurs de confusion pour corriger les tests d'association. Les méthodes statistiques développées lors de cette thèse reposent sur des problèmes de factorisation matricielle. Nous allons maintenant introduire plus en détails les problématiques biologiques.

1.1 La génomique des populations

La génétique des populations est l'étude de la distribution et des changements de la fréquence des variants d'un gène, que l'on appelle allèles, dans des populations d'êtres vivants. La génomique des populations est un néologisme qui souligne le fait que l'on n'étudie plus les différences génétiques dans et entre les populations à l'échelle d'un seul gène, mais à l'échelle du génome complet. La génomique des populations a des applications en épidémiologie où elle permet de comprendre la transmission des maladies génétiques. Elle est aussi utilisée en agronomie où des programmes de sélection modifient le patrimoine génétique de certains organismes pour créer des races ou variétés plus performantes, ou plus résistantes à des maladies. Elle permet également de comprendre les mécanismes de conservation et de disparition des populations et des espèces.

Les populations étudiées par la génétique des populations sont constituées d'un ensemble d'individus qui forme une unité de reproduction. Les individus d'une population peuvent se croiser entre eux, ils se reproduisent moins avec les individus des populations voisines, desquelles ils sont géographiquement isolés. Les changements de fréquence d'allèles dans les populations peuvent être expliqués par quatre pressions évolutives : la mutation, la sélection, la dérive génétique, et la migration. La mutation crée de nouveaux allèles qui peuvent être neutres, c'est-à-dire sans effet sur l'organisme. Les mutations non neutres peuvent par exemple avoir un effet sur la capacité de

l'organisme à s'adapter à son environnement. La sélection est l'hypothèse centrale de la théorie de Darwin. Si un allèle permet à l'organisme d'être mieux adapté à son environnement, alors il aura plus de chance d'être transmis aux générations futures. Par exemple, un organisme porteur d'un allèle apportant un avantage sélectif peut se reproduire plus facilement. La dérive génétique provient du fait qu'il n'existe pas de populations infinies. Ainsi, au fil des générations les distributions alléliques changent à cause de la reproduction au hasard des individus. Ce processus peut engendrer la disparition ou la fixation d'un allèle neutre simplement par hasard. Enfin, la migration est le passage de gènes d'une population à une autre (par des individus migrant d'une population à l'autre, par exemple sous forme de graines ou de pollen chez les plantes). Si cet échange se fait entre des populations ayant des fréquences alléliques différentes, il va tendre à modifier les fréquences alléliques.

La génétique des populations trouve ses origines dans les travaux de Sewall Wright, J. B. S. Haldane et Ronald Fisher. Bien que ces travaux soient antérieurs à la découverte de la structure de l'ADN (acide désoxyribonucléique) par Watson et Crick en 1953, la génomique des populations utilise des données d'observation de génomes de plusieurs individus provenant de plusieurs populations. L'ADN contient toute l'information génétique (le génome) permettant le développement, le fonctionnement et la reproduction des êtres vivants. L'ADN peut être considéré comme une longue séquence des nucléotides : A, C, T, G. Les récentes améliorations en séquençage de l'ADN ont permis d'acquérir les génomes de nombreuses espèces différentes. Dans le cadre de cette thèse, nous nous sommes intéressés seulement aux données dites de SNPs (single-nucleotide polymorphism), qui sont les polymorphismes génétiques d'un seul nucléotide (Figure 1.1). De plus, nous supposons que l'on peut seulement observer deux allèles par SNP. Cette hypothèse n'est pas réductrice car la probabilité qu'une mutation survienne deux fois à la même position est très faible, et les mutations sont des évènements rares pour les espèces considérées dans cette thèse [1]. Les SNPs représentent 90% de l'ensemble des variations génétiques humaines, et des SNPs avec une fréquence allélique supérieure à 1% sont présents dans le génome humain, en moyenne tous les cent à trois cents nucléotides. Nous appelons locus une position sur l'ADN ; nous parlons ainsi du locus d'un SNP. Nous représentons les données génétiques comme la matrice contenant le nombre de fois que l'allèle muté a été observé

1. Le taux de mutation chez l'humain est environ 0.5×10^{-9}.

pour chaque individu et chaque locus (Figure 1.2). Nous noterons $\mathbf{Y}$ la matrice de SNPs dans ce qui suit.

$$\text{ADNs} \left\{ \begin{array}{cccccccc} \cdots & \text{G} & \text{A} & \text{T} & \text{C} & \text{C} & \cdots & \cdots \\ \cdots & \text{G} & \text{A} & \text{A} & \text{C} & \text{C} & \cdots & \cdots \\ \cdots & \text{G} & \text{A} & \text{A} & \text{C} & \text{C} & \cdots & \cdots \\ \cdots & \text{G} & \text{A} & \text{T} & \text{C} & \text{C} & \cdots & \cdots \\ \cdots & \text{G} & \text{A} & \text{T} & \text{C} & \text{C} & \cdots & \cdots \end{array} \right.$$

FIGURE 1.1 – **Illustration d'un SNP.** Le nucléotide différent entre les séquences est un SNP.

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 2 & 2 & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ 0 & 0 & 2 & 0 & \cdots & \cdots & \cdots \end{bmatrix}$$

FIGURE 1.2 – **Illustration d'une matrice de SNPs pour une espèce diploïde.** Chaque élément de la matrice est le nombre de fois que l'allèle muté est observé pour un individu donné à un locus donné.

1.1.1 Estimation de la structure génétique des populations

Une étape très importante en génétique des populations est l'inférence d'une représentation synthétique de la structure de populations à partir des données génétiques. La structure de populations influence les distributions des SNPs par le biais des quatre pressions évolutives dont nous avons parlées dans le début de ce chapitre. Les pressions évolutives induisent une différenciation des distributions alléliques entre les populations. Nous avons illustré ce résultat en représentant les distributions alléliques d'un SNP pour des individus humains provenant de populations africaine, européenne et afro-américaine. Nous constatons une différence de distribution allélique entre les populations (Figure 1.3).

Une méthode très utilisée pour visualiser la structure de population est l'analyse en composantes principales (ACP). En effet, dans une population d'individus structurée en K populations, il faut $K - 1$ axes principaux pour représenter la structure de populations à partir de données génétiques (PATTERSON et al., 2006). Nous proposons d'illustrer ce résultat en calculant les deux premiers axes principaux d'un échantillon

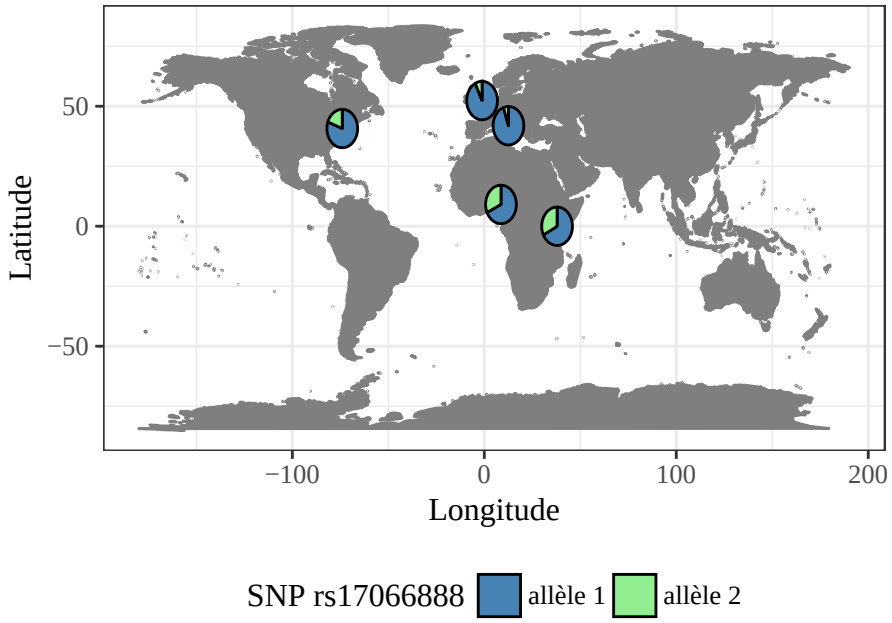

FIGURE 1.3 – **Différenciation allélique entre des populations**. Distribution des allèles du SNP rs17066888 dans des populations européenne, africaine et afro-américaine.

de données de SNPs composé d'individus humains de populations africaine, européenne et afro-américaine. Les deux premiers axes principaux permettent de visualiser un groupe composé des individus européens et deux groupes composés des individus africains. Les individus afro-américains sont répartis entre les groupes européens et afro-américains (Figure 1.4)

Un modèle très utilisé pour étudier la structure génétique des populations à partir de données de SNPs est celui du logiciel `structure` (PRITCHARD et al., 2000). Dans ce modèle, nous supposons que le génome de chaque individu est la combinaison de morceaux de génomes provenant de K groupes génétiques, aussi appelés populations ancestrales. Dans le cadre de ce modèle, nous pouvons écrire

$$\Pr(\mathbf{Y}_{i,\ell} = j) = \sum_{k=1}^{K} \mathbf{G}_{(d+1)\ell+j,k}\mathbf{Q}_{i,k}, \tag{1.1}$$

où $\Pr(\mathbf{Y}_{i,\ell} = j)$ est la probabilité d'observer l'allèle j au locus ℓ chez l'individu i. Le terme $\mathbf{G}_{(d+1)\ell+j,k}$ représente la fréquence d'apparition de l'allèle j au locus ℓ dans le groupe génétique k. Le terme $\mathbf{Q}_{i,k}$ (appelé coefficient de métissage ou d'ascendance) est la proportion de gènes de l'individu i provenant du groupe k. Les coefficients

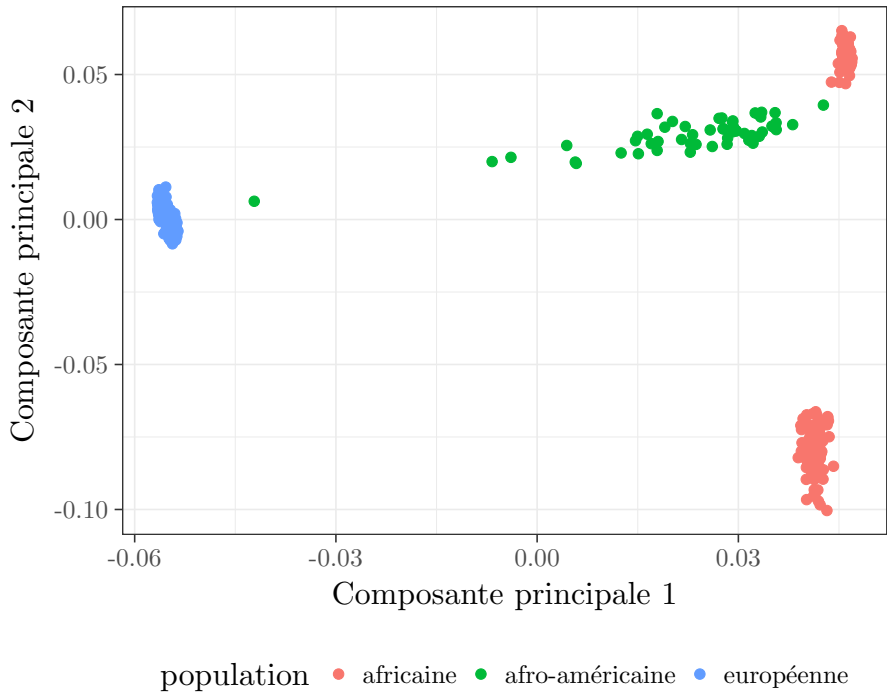

FIGURE 1.4 – **Visualisation de la structure de population avec l'ACP.** Scores des deux premières composantes principales calculées sur des données de SNPs d'invidus humains de populations européenne, africaine et afro-américaine.

d'ascendance et les fréquences d'allèles dans les groupes génétiques sont respectivement rangés dans des matrices **Q** et **G**. Afin d'illustrer le modèle de `structure`, nous avons calculé les coefficients de métissage sur le jeu de données utilisé précédemment pour illustrer l'ACP. Nous avons utilisé le logiciel `snmf` qui permet de calculer des coefficients de métissage à partir de données de SNPs avec $K = 2$ groupes génétiques (FRICHOT et FRANÇOIS, 2015). Les groupes génétiques trouvés par le logiciel `snmf` sont européen et africain ; tandis que les individus afro-américains ont des génomes provenant des groupes génétiques africain et européen (Figure 1.5). Il s'agit du résultat attendu au regard de l'histoire démographique des individus afro-américains (TISHKOFF et al., 2009).

1.1.2 Méthodes d'estimation des coefficients d'ascendance

Il existe de nombreuses méthodes pour estimer les coefficients d'ascendance à partir de données génétiques. Le modèle du logiciel `structure` est bayésien et l'inférence repose sur des méthodes d'échantillonnage de la loi a posteriori des coefficients d'as-

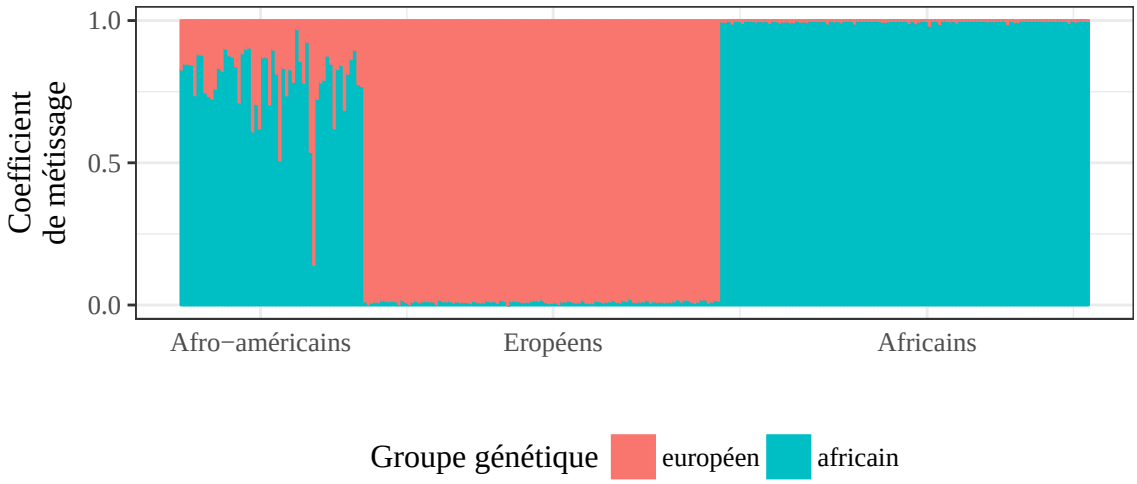

FIGURE 1.5 – **Coefficients de métissage.** Estimation par le logiciel `snmf` des coefficients de métissage pour un jeu de données composé d'individus humains provenant de populations européenne, africaine et afro-américaine.

cendance (PRITCHARD et al., 2000). D'autres méthodes visant à rendre plus rapide l'inférence des matrices d'ascendance, minimisent la fonction log-vraisemblance des paramètres d'ascendance génétique (TANG et al., 2005 ; ALEXANDER, NOVEMBRE et al., 2009). Certaines méthodes utilisent des méthodes d'inférence variationnelle bayésiennes (RAJ et al., 2014). Des méthodes très rapides, ne reposant pas sur une modélisation probabiliste, ont aussi été proposées pour passer à l'échelle des grands jeux de données modernes (FRICHOT, MATHIEU et al., 2014 ; POPESCU et al., 2014). Par ailleurs, de nombreuses méthodes utilisent l'information spatiale individuelle afin d'améliorer l'estimation de l'ascendance génétique et de localiser les groupes génétiques dans l'espace. Des méthodes ont ajouté l'information géographique au modèle bayésien de `structure` (C. CHEN et al., 2007 ; CORANDER et al., 2008 ; GUEDJ et GUILLOT, 2011). Cependant, les méthodes bayésiennes reposent sur de nombreuses hypothèses et passent plus difficilement à l'échelle des grands jeux de données. Aucune méthode non basée sur un modèle probabiliste n'a été proposée pour l'inférence spatiale des coefficients d'ascendance. Nous résumons les méthodes d'inférence des coefficients d'ascendance génétique dans la Table 1.1

TABLE 1.1 – **Méthodes spatiales et non spatiales d'estimation de coefficients d'ascendance.** Les méthodes TESS3-AQP/APLS sont présentées dans cette thèse.

Méthode	Modèle	Spatiale	Algorithme	Référence
STRUCTURE	bayésien	non	MCMC	PRITCHARD et al. (2000) et FALUSH et al. (2003)
FRAPPE	vraisemblance	non	EM	TANG et al. (2005)
TESS	bayésien	oui	MCMC	C. CHEN et al. (2007)
GENELAND	bayésien	oui	MCMC	GUEDJ et GUILLOT (2011)
BAPS	bayésien	oui	optimisation stochastique	CORANDER et al. (2008)
ADMIXTURE	vraisemblance	non	optimisation quasi-Newton alternée	ALEXANDER, NOVEMBRE et al. (2009) et ALEXANDER et LANGE (2011)
fastStructure	bayésien	non	inférence variationnelle bayésienne	RAJ et al. (2014)
PSIKO	ACP	non	SVD	POPESCU et al. (2014)
sNMF	factorisation matricielle parcimonieuse	non	optimisation quadratique alternée avec projection	FRICHOT, MATHIEU et al. (2014)
TESS3-AQP	factorisation matricielle régularisée sur graphe	oui	optimisation quadratique alternée	
TESS3-APLS	factorisation matricielle régularisée sur graphe	oui	moindres carrés alternés projetés	
conStruct	bayésien	oui	MCMC	BRADBURD et al. (2017)

1.2 Tests d'association

Un problème fondamental en science du vivant consiste à détecter les relations de causalité qui existent entre des événements. En statistique, un événement est modélisé par une variable aléatoire. Il est seulement possible de détecter des liens statistiques entre les variables aléatoires ; on parle alors d'étude de corrélations. La corrélation renseigne sur les probabilités jointes des variables aléatoires en question. Dans un cadre statistique, nous parlons de tests d'association lorsque l'on cherche à identifier des corrélations entre des variables aléatoires.

Les tests d'association sont très utilisés en génétique pour comprendre les fonctions des gènes. Par exemple, on peut chercher quels SNPs sont corrélés à une maladie pour comprendre les causes génétiques de celle-ci. Cependant, comme nous l'avons vu dans la partie précédente, il existe de nombreux facteurs responsables de la diversité génétique. Quand les facteurs de variation du génome sont corrélés à la variable d'étude (la maladie par exemple), alors les études d'association sont faussées. On observe en général une augmentation du nombre de gènes associés à la variable d'étude. Une telle situation n'est pas souhaitable car bien qu'il s'agisse de corrélations, il ne s'agit pas de corrélations intéressantes pour l'étude biologique. Nous expliquons maintenant plus en détail ce que sont les facteurs de confusion dans les études d'association.

1.2.1 Les facteurs de confusion

Basées sur l'analyse de la corrélation, les études d'association sont confrontées aux problèmes des facteurs de confusion et de la causalité. En effet, lorsque l'on détecte une corrélation entre deux variables, cela n'implique pas nécessairement qu'il y a un lien de causalité entre celles-ci. Le lien de causalité entre les deux variables peut être bien plus complexe, et notamment impliquer des liens avec d'autres variables non observées. En particulier, il est possible de conclure à une association entre deux variables alors qu'elles sont associées à une autre variable non considérée dans l'étude. On appelle alors la variable non observée un facteur de confusion. La figure 1.6 illustre cette situation. Le problème des facteurs de confusion est connu depuis longtemps. En effet, on le retrouve déjà dans l'ouvrage *The Design of Experiment* de Ronald Fisher qui introduisit entre autre le concept d'hypothèse nulle en statistique (Fisher, 1937). Dans cette thèse nous nous intéressons aux études d'association à très

grande échelle. Nous avons d'une part des observations de p variables sur n individus rassemblées dans une matrice $\mathbf{Y}$ de taille $n \times p$, et en général p est très grand devant n. Nous avons d'autre part l'observation d'une variable sur les mêmes n individus que l'on rassemble dans une matrice $\mathbf{X}$, de taille $n \times 1$. L'objectif est alors de trouver parmi les p variables $\mathbf{Y}$ celles qui sont associées à $\mathbf{X}$. Nous supposons de plus qu'il existe un certain nombre de variables non observées qui permettent d'expliquer les variations de $\mathbf{Y}$. Ces variables non observées, que l'on appellera variables latentes, sont potentiellement des facteurs de confusion pour l'étude d'association entre les matrices $\mathbf{Y}$ et $\mathbf{X}$. Les variables latentes sont potentiellement corrélées à $\mathbf{X}$; il faut donc les prendre en compte dans l'étude d'association.

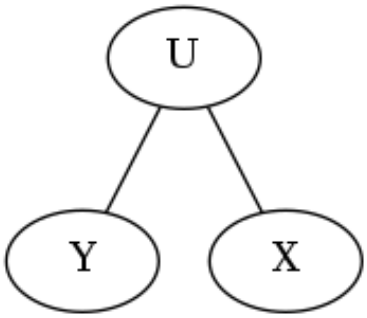

FIGURE 1.6 – **Graphe de corrélation entre la variable Y, la variable X et le facteur de confusion U.** Dans cette situation si on ne prend pas en compte la variable **U** dans l'étude d'association alors **X** et **Y** apparaîtront comme étant associées.

1.2.2 Simulation numérique d'une association avec facteurs de confusion

Dans cette partie nous proposons de montrer l'intérêt de prendre en considération les facteurs de confusion dans les études d'association par une simulation numérique. Pour cela nous simulons une variable explicative $\mathbf{X}$ et une variable latente $\mathbf{U}$, de sorte que le coefficient de corrélation entre les deux variables soit égal à 0.6. Nous simulons ensuite une matrice de bruit gaussien de moyenne nulle et variance égale à 1, notée $\mathbf{E}$. La matrice des effets de la variable latente sur $\mathbf{Y}$ est aussi simulée à l'aide de la loi normale. Nous notons la matrice des effets latents $\mathbf{V}$. La matrice des effets de $\mathbf{X}$ sur $\mathbf{Y}$, notée $\mathbf{B}$, est simulée de sorte que 1% de ses lignes soient non nulles. Enfin, la

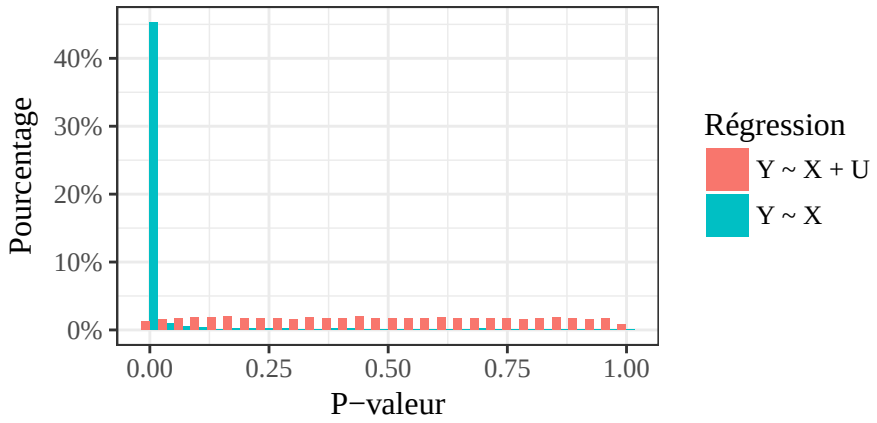

FIGURE 1.7 – **Test de nullité des coefficients de la régression sans et avec le facteur de confusion.** Les données ont été simulées avec une variable latentes **U** corrélée avec la variable **X**.

matrice des variables expliquées, **Y**, est calculée telle que

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^T + \mathbf{X}\mathbf{B}^T + \mathbf{E}. \tag{1.2}$$

Cette simulation correspond à une situation où 1% des colonnes de **Y** sont associées à **X**. La variable latente **U** est un facteur de confusion pour cette étude d'association car elle est corrélée à la variable **X**.

Afin de détecter les variables expliquées associées à la variable explicative, nous réalisons une régression linéaire de **Y** par **X**. Nous effectuons une seconde régression linéaire avec cette fois la variable **X**, ainsi que la variable latente **U**, comme variables explicatives de la régression. Nous réalisons un test de Student pour tester la nullité des coefficients associés à la variable **X** dans chacune des deux régressions. Quand on ne prend pas en compte la variable latente, plus de 40% des p-valeurs sont inférieures à 10^{-15} ; alors que quand on prend en compte les facteurs latents, la distribution des p-valeurs est bien uniforme comme on s'y attend (Figure 1.7). En effet, on s'attend à une distribution uniforme des p-valeurs car la majorité des colonnes de **Y** ne sont pas associées à la variable **X** (seulement 1% y sont associées par simulation). Dans le cas de cette simulation il est impossible de ne pas prendre en compte la variable latente ; sans celle-ci on détecte presque la moitié des colonnes de **Y** comme étant associées à **X**.

1.2.3 Méthodes de correction des facteurs de confusion pour les études d'association

Prendre en compte les facteurs latents est un problème important des études d'association. Certaines méthodes utilisent l'analyse en composantes principales pour estimer les facteurs de confusion et les intégrer aux tests de significativité (Rahmani et al., 2016 ; Price et al., 2006). D'autres méthodes utilisent les modèles mixtes afin de corriger les tests d'hypothèse pour les sources de variation indésirable (Kang et al., 2008 ; X. Zhou et Stephens, 2014 ; Loh et al., 2017). Récemment, de nombreuses méthodes ont été proposées pour permettre d'estimer dans un même modèle les effets des variables latentes et les effets des variables étudiées pour l'association. La plupart de ces méthodes reposent sur l'équation (1.2) que nous avons utilisée dans la partie précédente. Ces modèles ont reçu plusieurs noms dans la littérature : Latent Fator Mixed Models (LFMM) (Frichot, Schoville et al., 2013), regression-based latent models (RLFM) (Agarwal et B.-C. Chen, 2009), factor-augmented regression model (Gerard et Stephens, 2017a), surrogate variable analysis (SVA) (Leek et Storey, 2007). Nous résumons dans le Table 1.2, les méthodes reposant sur des modèles construits à partir de l'équation (1.2) .

1.3 Résumé de la problématique

L'évaluation de l'ascendance génétique est un problème majeur en génétique des populations. Des méthodes très efficaces ont été proposées pour l'estimation des coefficients d'ascendance à partir de données génétiques. Cependant, l'estimation de l'ascendance génétique en intégrant l'information spatiale a reçu moins d'attention. Ainsi, afin de permettre l'étude des données génétiques modernes, le développement de méthodes statistiques pour l'inférence de l'ascendance génétique intégrant l'information spatiale est nécessaire.

Répondant à l'arrivée massive de données, le développement de méthodes pour les études d'association à grandes échelles, est très actif en ce moment. Dans les études d'association, un aspect très important est de détecter et corriger les sources de variation indésirable pour l'étude. Chaque méthode utilise des approches différentes et aucune ne s'est imposées comme étant la méthode de référence. Dans le contexte actuel il est nécessaire de développer des méthodes rapides et utilisables sur les données

TABLE 1.2 – **Méthodes reposant sur l'équation 1.2 pour la correction des facteurs de confusion dans les études d'association.** Les méthodes ridgeLFMM et LassoLFMM sont présentées dans cette thèse. Les méthodes cate, sva-twostep et sva-riw sont présentées plus en détail dans la section 3.4.

Méthode	Modèle	Algorithme	Test d'hypothèse	Référence
sva-twostep	ACP et régression linéaire	moindres carrés ordinaire et SVD	test de Fisher	LEEK et STOREY (2007)
sva-irw	*weighted*-ACP et régression linéaire	moindres carrés ordinaire et *weighted*-SVD	test de Fisher	LEEK et STOREY (2008)
RLFM	bayésien	Monte-Carlo EM	pas de test	AGARWAL et B.-C. CHEN (2009)
famt	vraisemblance	EM	test de Student	FRIGUET et al. (2009)
LFMM	bayésien	MCMC	test de wald, estimation de la variance par bootstrap bayésien	FRICHOT, SCHOVILLE et al. (2013)
cate	analyse factorielle et régression linéaire	EM ou SVD et estimation des moindres carrés généralisée	test basé sur la distribution asymptotique de l'estimateur des effets d'intérêt	WANG et al. (2017)
ridgeLFMM	factorisation matricielle avec régularisation L_2	SVD et estimation des moindres carrés régularisée en norme L_2	test de Student	
lassoLFMM	factorisation matricielle avec régularisation L_1	*soft-thresholded* SVD et estimation des moindres carrés régularisée en norme L_1	test de Student	
MOUTHWASH / BACKWASH	régression linéaire et analyse factorielle	moindres carrés ordinaire et EM ou descente par coordonnées	*adaptive shrinkage* (ASH) (STEPHENS, 2016)	GERARD et STEPHENS (2017a)

massives. Par ailleurs, une comparaison des méthodes de correction pour les facteurs de confusion permettrait de mieux comprendre les spécificités de chaque méthode.

1.4 Contexte de la thèse

Cette thèse a été financée par le LabEx PERSYVAL-Lab et co-encadrée par Olivier François du laboratoire TIMC-IMAG et Olivier Michel du laboratoire GIPSA-lab. Nos travaux ont principalement été réalisés au sein de l'équipe BCM (Biologie Computationnelle et Mathématique). L'équipe BCM du laboratoire TIMC-IMAG est spécialisée en étude de données génétiques et en développement de modèles mathématiques pour les systèmes biologiques complexes. Les méthodes présentées dans cette thèse sont donc dans la lignée des méthodes développées au sein de l'équipe BCM. Ainsi, les logiciels que nous avons développés dans cette thèse viennent compléter les logiciels produits par l'équipe BCM. Nous parlons en particulier de `TESS` 2.3, un logiciel d'inférence spatiale de la structure génétique des populations (Durand et al., 2009) ; ainsi que de `LEA`, un package R proposant une suite de fonctions dédiées aux études d'association génomique (Frichot et François, 2015).

1.5 Objectifs de la thèse

La génétique produit beaucoup de données grâce aux technologies de séquençage toujours plus efficaces. Cette affluence de données pose de nouveaux problèmes aux statisticiens. L'objectif de cette thèse de doctorat est d'améliorer les outils statistiques qui permettent aux biologistes de répondre à des questions concrètes sur le vivant. Dans cette thèse nous nous sommes intéressés à deux problématiques en analyse de données génétiques : l'estimation de la structure génétique de population et les études d'association. L'accent a été mis sur la complexité des algorithmes développés afin qu'ils soient applicables aux données génétiques modernes. De plus, l'importance a été placée aussi bien sur le développement mathématique des méthodes que sur leur implémentation informatique. Afin que nos méthodes statistiques soient utilisables par la communauté scientifique, il a été important de rendre accessibles des implémentations informatiques efficaces de nos nouvelles méthodes statistiques.

1.6 Résumé des résultats principaux

Dans le cadre de cette thèse, nous avons proposé plusieurs méthodes reposant sur des problèmes de factorisation matricielle. Nos méthodes ont été implémentés dans deux packages R : `tess3r` et `lfmm`. Le package `tess3r` contient les algorithmes AQP et APLS d'inférence de l'ascendance génétique en incluant l'information spatiale. Le package `lfmm` contient les algorithmes lassoLFMM et ridgeLFMM d'estimation des facteurs de confusion pour corriger les études d'association. Nous détaillons maintenant les résultats pour chacun des deux logiciels.

1.6.1 `tess3r`

Dans le package `tess3r`, nous avons développé des algorithmes d'estimation rapide des coefficients de métissage à partir de données génétiques et géographiques. Les algorithmes reposent sur un problème de factorisation de la matrice génétique. L'objectif est de factoriser la matrice génétique en le produit d'une matrice des coefficients d'ascendance et d'une une matrice des fréquences d'allèle dans les groupes génétiques. Pour inférer les matrices d'ascendance nous avons utilisé une approximation des moindres carrés. L'information spatiale est ajoutée à la fonction objectif au moyen d'une régularisation sur la matrice des coefficients de métissage. La régularisation spatiale permet d'ajouter l'hypothèse que des individus proches ont plus de chance de partager des ancêtres communs, que des individus éloignés. Afin d'estimer les matrices d'ascendance, nous avons proposé deux algorithmes appelés AQP et APLS. Nos algorithmes diffèrent dans les approximations qu'ils font pour diminuer la complexité algorithmique. Plus précisément, nous avons d'une part l'algorithme AQP qui alterne des résolutions de problèmes d'optimisation quadratique. Le corollaire 2 établi par Grippo et Sciandrone (2000) permet de montrer la convergence de l'algorithme AQP vers un minimum local de la fonction objectif. Nous avons d'autre part, l'algorithme APLS pour lequel nous avons supprimé les contraintes des problèmes d'optimisation quadratique. Cela permet d'alterner la résolution de problèmes des moindres carrés régularisés par une norme L_2. Ainsi, la complexité de l'algorithme APLS augmente linéairement avec le nombre d'individus dans l'échantillon. Enfin, nous avons mis en place un test de détection de l'adaptation locale. La statistique de test est calculée à partir des estimations spatiales des matrices d'ascendance génétique

(Martins et al., 2016).

En utilisant des simulations de coalescents, nous avons montré que les deux algorithmiques AQP et APLS retournent des résultats avec la même précision statistique. Toujours sur des simulations de coalescents, nous avons montré que nos algorithmes reproduisent les mêmes erreurs statistiques que le logiciel `TESS` 2.3 (C. Chen et al., 2007). Le logiciel `TESS` 2.3 permet également l'estimation des coefficients de métissage à partir de données génétiques et géographiques mais en utilisant un modèle bayésien. Sur ces simulations, nos algorithmes étaient 10 à 100 fois plus rapides que le logiciel `TESS` 2.3.

Pour mesurer le bénéfice de l'utilisation d'algorithmes spatiaux, nous avons comparé les erreurs statistiques observées pour les algorithmes spatiaux avec celles observées pour un algorithme non spatial `snmf` (Frichot, Mathieu et al., 2014). Dans nos expériences numériques, les erreurs des méthodes spatiales sont inférieures à celles observées avec des méthodes non spatiales. De plus, les algorithmes spatiaux ont permis de détecter une structure de population plus subtile.

Enfin, nous avons illustré l'utilisation de notre package R sur un millier de génotypes *A.thaliana*, chacun incluant plus de 210k SNPs. Notre méthode a permis d'exiber la structure de population de l'espèce *A.thaliana* en Europe. Par ailleurs, nous avons appliqué les tests de neutralité afin d'effectuer un balayage du génome pour la sélection dans des écotypes européens de l'espèce végétale *A.thaliana*. Le scan du génome a confirmé la preuve de la sélection des gènes liés à la floraison *CIP4.1*, *FRI* et *DOG1* différenciant la Fenno-Scandinavie du nord-ouest de l'Europe (Horton et al., 2012).

1.6.2 `lfmm`

Dans le package `lfmm`, nous avons développé des algorithmes qui permettent d'estimer les variables latentes afin de corriger les études d'association pour les facteurs de confusion. Nous avons proposé une fonction objectif basée sur l'approximation des moindres carrés de l'égalité (1.2) du modèle mixte à facteurs latents. Le modèle consiste à expliquer les variations des variables étudiées par la somme de deux effets : l'effet des variables latentes et l'effet des variables explicatives (ces dernières sont aussi appelées covariables). L'attache aux données de la fonction objectif a été construite à partir de l'approximation des moindres carrés de l'égalité (1.2). Nous avons montré

que le terme d'attache aux données ne permet pas à lui seul de définir des estimateurs de manière univoque pour notre problème. Ainsi, nous avons proposé d'ajouter un terme de régularisation portant sur les effets des variables explicatives. Nous avons appelé nos méthodes ridgeLFMM, pour la régularisation L_2 et lassoLFMM, pour la régularisation L_1. L'algorithme ridgeLFMM utilise la formule du minimum global de la fonction objectif des moindres carrés régularisée en norme L_2. Nous avons apporté la démonstration de cette formule. Pour l'algorithme lassoLFMM, nous avons proposé une méthode alternée de descente par blocs de coordonnées. Les travaux de Tseng (2001) permettent de démontrer la congergence de l'algorithme lassoLFMM vers le minimum global de sa fonction objectif. Cependant, nous avons proposé une preuve adaptée au résultat de convergence de notre algorithme.

Pour évaluer la capacité de nos méthodes à corriger les études d'association pour les facteurs de confusion, nous avons réalisé des simulations à partir d'un jeu de données réelles de génotypes humains. Nous avons ajouté à la comparaison une méthode de référence qui ne prend pas en compte les facteurs latents. Nous avons aussi comparé les méthodes de la littérature reposant sur le même modèle de régression avec facteur latent que nos méthodes : cate, sva-irw et sva-twostep. Nous avons également considéré la méthode calculant les facteurs latents à l'aide de l'ACP. Ces simulations nous ont permis de montrer que nos méthodes ridgeLFMM et lassoLFMM ont la même puissance que la méthode oracle qui connait les variables latentes de la simulation. De plus, les statistiques p-valeurs obtenues avec nos méthodes sont correctement calibrées. Nous avons observé que la méthode cate obtenait des performances très proches de celles de nos méthodes sur toutes les simulations considérées.

Enfin, nous avons illustré l'utilisation de nos méthodes sur des études d'association pour des données réelles. Sur les données réelles, nous montrons que nos méthodes permettent de retrouver les associations découvertes par d'autres études. De plus, nous observons dans ces études que malgré les ressemblances conceptuelles entre les méthodes, les associations découvertes peuvent varier largement d'une méthode à l'autre. Cela met en avant la nécessité d'utiliser plusieurs méthodes dans les études d'association, ainsi que d'être prudent dans les interprétations.

Chapitre 2

Inférence des coefficients de métissage à l'aide de données géographiques

2.1 Résumé

L'évaluation précise de la répartition de l'ascendance génétique dans l'espace géographique est l'une des principales questions abordées par les biologistes de l'évolution. Cette question a été communément abordée par l'application de programmes d'estimation bayésiens permettant à leurs utilisateurs d'estimer les proportions individuelles de métissage et les fréquences alléliques parmi les populations ancestrales putatives. Suite à l'explosion des technologies de séquençage à haut débit, plusieurs algorithmes ont été proposés pour faire face au fardeau de calcul généré par les données massives dans ces études. Dans ce contexte, l'intégration de la proximité géographique dans les algorithmes d'estimation de l'ascendance est un défi statistique et computationnel ouvert. Dans ce chapitre, nous introduisons de nouveaux algorithmes qui utilisent l'information géographique pour estimer les proportions d'ascendance et les fréquences génotypiques ancestrales à partir des données génétiques de la population étudiée. Nos algorithmes combinent les méthodes de factorisation matricielle et les statistiques spatiales pour fournir des estimations des matrices d'ascendance basées sur l'approximation des moindres carrés. Nous démontrons le bénéfice de l'utilisation d'algorithmes spatiaux grâce à des simulations numériques, et nous fournissons un exemple d'application de nos nouveaux algorithmes à un ensemble d'échantillons référencés spatialement pour les espèces végétales *Arabidopsis thaliana*. Sans perte de précision statistique, les nouveaux algorithmes présentent des temps d'exécution beaucoup plus courts que ceux observés pour les méthodes spatiales développées antérieurement. Nos algorithmes

sont implémentés dans le package R, `tess3r`.

2.2 Introduction

Représenter la structure génétique de population est une étape importante dans l'étude de données génétiques. Les données génétiques sont volumineuses et multivariées. La structure génétique des populations fournit une représentation synthétique qui permet de visualiser la variation génétique induite par la stratification en populations. La stratification en populations fournie des informations sur l'histoire et l'évolution démographique de l'espèce étudié (J. Z. Li et al., 2008). Il est également indispensable de l'utiliser comme facteur de correction dans les études d'association à un phénotype, un gradient environnemental ou encore une maladie (Marchini et al., 2004). De même, il existe de nombreuses applications en médecine génétique nécessitant de connaître la structure de populations, comme par exemple le calcul d'un score de risque génétique pour une maladie (Wray et al., 2013). Enfin, l'étude de la répartition en population d'une espèce dans son habitat est une étape clé en génétique du paysage (François et Waits, 2015).

Pour modéliser la structure génétique des populations, nous supposons que le génome de chaque individu est la combinaison de morceaux de génomes provenant de K groupes génétiques ; les groupes génétiques sont aussi appelés populations ancestrales (Pritchard et al., 2000). Dans chaque groupe génétique, l'objectif est d'estimer les fréquences d'allèle pour chaque SNPs. Pour chaque individu, il faut estimer la proportion de son génotype qui provient de chaque groupe génétique. Les proportions sont appelées coefficients de métissage, aussi appelés coefficients d'ascendance.

2.2.1 Méthodes d'inférence des coefficients de métissage

L'inférence des coefficients de métissage a été largement étudiée et il existe de nombreuses méthodes. On distingue deux types d'approche : les approches reposant sur un modèle probabiliste et les approches fondées sur l'optimisation d'une fonction objectif.

Parmi les approches reposant sur un modèle probabiliste, on compte le logiciel `structure` proposé par Pritchard et al. (2000) qui a introduit le modèle de structure génétique de population dont nous avons parlé dans l'introduction. L'accès à

des données génétiques de plus en plus massives a provoqué l'émergence de plusieurs algorithmes plus rapides que celui de `structure`. En effet, le logiciel `structure` implémente un algorithme d'échantillonnage de Monte-Carlo pour estimer la distribution a posteriori des coefficients de métissage et des fréquences d'allèle dans les groupes génétiques. Cependant les algorithmes de Monte-Carlo ne passent pas à l'échelle des grands jeux de données génétiques modernes. Il a été proposé des améliorations du logiciel `structure` reposant sur une fonction de vraisemblance définie pour la matrice des coefficients de métissage et les fréquences d'allèle dans les groupes. L'estimation est effectuée en maximisant la fonction log-vraisemblance. Une première amélioration de l'algorithme `structure` est fondée sur un algorithme EM (Expectation Maximisation) maximisant la fonction de vraisemblance (Tang et al., 2005). Des algorithmes de vraisemblance plus récents sont implémentés dans les programmes `admixture` et `fastStructure` (Alexander et Lange, 2011 ; Raj et al., 2014).

Dans les approches reposant sur l'optimisation d'une fonction objectif, les coefficients de métissage sont estimés à l'aide de méthodes de moindres carrés ou d'analyse factorielle. Pour estimer les matrices des coefficients de métissage et de fréquence d'allèle dans les groupes, Engelhardt et Stephens (2010) proposent d'utiliser une analyse parcimonieuse à facteurs ; Frichot, Mathieu et al. (2014) utilisent des algorithmes de factorisation de matrice non négative ; Popescu et al. (2014) utilisent l'analyse en composantes principales. Ces méthodes, reposant sur des problèmes d'optimisation, permettent de reproduire avec précision les résultats des approches considérant une fonction de vraisemblance (Frichot, Mathieu et al., 2014). En outre, cette catégorie de méthodes fournit des algorithmes qui sont généralement plus rapides que ceux des méthodes reposant sur un modèle probabiliste.

2.2.2 Méthodes d'inférence des coefficients de métissage à l'aide de données géographiques

Dans la nature, les individus d'une espèce évoluent dans un environnement géographique. Les groupes génétiques, identifiés par les méthodes d'estimation de la structure génétique des populations, sont induits par les pressions évolutives qui s'opèrent dans l'environnement géographique de l'espèce. Les groupes génétiques peuvent par exemple être générés par l'isolation des populations à cause d'une mer les séparant ou bien des différences d'altitude entre celles-ci. L'étude réalisée par Novembre et al. (2008) a

montré qu'il est possible de prédire la position des individus à partir de l'étude de la structure génétique des populations. De nombreuse méthodes ont permis d'améliorer la prédiction de la position géographique des individus à partir du génome (Baran et al., 2013 ; Yang et al., 2012 ; Bhaskar et al., 2016 ; Rañola et al., 2014). Si la structure génétique des populations permet de prédire la position spatiale des individus, alors il est possible d'améliorer l'estimation de la structure génétique des populations en utilisant l'information géographique. Cette idée a été exploitée pour améliorer le modèle bayésien de `structure` en intégrant des données géographiques dans la distribution a priori des coefficients de métissage (C. Chen et al., 2007 ; Corander et al., 2008). Les algorithmes spatiaux fournissent des estimations de la structure de population plus robustes que des algorithmes non spatiaux qui peuvent conduire à des estimations biaisées du nombre de groupes (Durand et al., 2009). Certaines méthodes bayésiennes sont basées sur des algorithmes de Monte-Carlo de chaîne de Markov qui nécessitent beaucoup de calcul (FRANÇOIS et DURAND, 2010). Ainsi, les méthodes existantes d'estimation des coefficients d'ascendance à l'aide de données géographiques ne sont pas adaptées aux grands jeux de données modernes.

2.2.3 Plan du chapitre

Dans ce chapitre, nous présentons une nouvelle méthode pour l'estimation des coefficients individuels de métissage à partir de données géographiques et génétiques. Cette méthode repose sur un problème de factorisation de matrices avec des contraintes convexes et une régularisation sur un graphe spatial. Nous proposons deux algorithmes qui résolvent le problème de factorisation. Le premier algorithme repose sur un algorithme d'optimisation quadratique alternée (AQP pour alternated quadratic programing), l'autre sur un algorithme des moindres carrés alternés projetés (APLS pour alternated projected least square). Le terme alterné dans les deux algorithmes fait référence au fait que l'on alterne une étape d'optimisation, selon la matrice des coefficients de métissage, puis selon la matrice des fréquences de génotypes ancestraux. L'algorithme AQP a un fondement théorique bien établi par Bertsekas (1997) ; ce n'est pas le cas de l'algorithme APLS. En utilisant des simulations coalescentes, nous montrons que les estimations calculées par l'algorithme APLS sont de bonnes approximations des solutions de l'algorithme AQP. De plus, nous montrons que les

performances de l'algorithme APLS s'élèvent aux dimensions des jeux de données modernes. Sur des simulations, nous montrons que l'erreur statistique fournis par APLS est du même ordre que l'erreur obtenue avec le logiciel `TESS` 2.3 ; ce logiciel implémente une méthode bayésienne pour l'inférence spatiale de la structure génétique des population. Toujours sur des simulations, nous montrons que notre algorithme spatial APLS estime mieux la structure de population que la méthode sNMF. L'algorithme de sNMF repose aussi sur un problème de factorisation de matrice mais n'utilise pas l'information spatiale. Enfin, nous présentons l'application de nos algorithmes aux données d'écotypes européens de l'espèce végétale *Arabidopsis thaliana*, pour lesquelles des données géographiques individuelles et génétiques sont disponibles (Horton et al., 2012).

2.3 Nouvelles méthodes d'estimation des coefficients de métissage

Dans cette section, nous présentons deux nouveaux algorithmes d'estimation de la structure génétique des populations qui intègrent l'information de proximité géographique.

2.3.1 Matrices d'ascendance génétique

Nous considérons une matrice de génotype, $\mathbf{Y}$, enregistrant des données de n individus à p locus polymorphes pour une espèce ayant une ploïdie de d, c'est-à-dire qui possède un génome composé de d exemplaires de chaque chromosome. Pour les SNPs autosomiques [1] dans un organisme diploïde, le génotype au locus ℓ est un nombre entier, 0, 1 ou 2, correspondant au nombre d'allèles de référence observé à ce locus. Dans nos algorithmes nous utilisons des formes disjonctives introduites par Frichot, Mathieu et al. (2014) pour coder les génotypes. Par exemple pour un organisme diploïde, le nombre d'allèles observés à chaque locus, 0, 1, 2, est encodé comme 100, 010 et 001. Pour les organismes d-ploïde, il existe $(d+1)$ génotypes possibles à chaque locus, et chaque valeur est encodée sous une forme disjonctive unique.

1. Les SNPs autosomiques sont les SNPs des chromosomes autosomes ou homologue. Les chromosomes autosomes sont formés de paires dont les membres ont la même forme, mais diffèrent des autres paires dans une cellule diploïde. Chez l'humain, on compte 22 paires de chromosomes homologues.

En utilisant la même approche que Frichot, Mathieu et al. (2014), si l'on suppose qu'il y a K groupes génétiques, nous cherchons à décomposer la matrice $\mathbf{Y}$ en une matrice de coefficients de métissage $\mathbf{Q}$, de taille $n \times K$ et une matrice de fréquences de génotypes dans les K groupes génétiques $\mathbf{G}$, de taille $(d+1)p \times K$. Nous notons $\mathbf{Q}_{i,k}$ le coefficient de métissage de l'individu i pour le groupe k. Nous avons de plus

$$\mathbf{Q} \geq 0\,, \quad \sum_{k=1}^{K} \mathbf{Q}_{i,k} = 1, \quad i = 1...n. \tag{2.1}$$

Nous notons $\mathbf{G}_{(d+1)\ell+j,k}$ la fréquence du génotype j au locus ℓ dans le groupe k et nous avons

$$\mathbf{G} \geq 0\,, \quad \sum_{j=0}^{d} \mathbf{G}_{(d+1)\ell+j,k} = 1, \quad \ell = 1...p. \tag{2.2}$$

Enfin, nous voulons estimer les matrices $\mathbf{Q}$ et $\mathbf{G}$ en factorisant la matrice de génotype de la façon suivante

$$\mathbf{Y} = \mathbf{Q}\mathbf{G}^T. \tag{2.3}$$

Ainsi le problème d'inférence peut être résolu en utilisant les méthodes de factorisation de matrices non négatives avec en plus les contraintes convexes décrites par les équations (2.1) et (2.2) (Lee et Seung, 1999 ; Cichocki et al., 2009). Dans la suite, nous utiliserons les notations Δ_Q et Δ_G pour représenter les ensembles formés à partir des contraintes sur les matrices $\mathbf{Q}$ et $\mathbf{G}$.

2.3.2 Information géographique

L'information géographique est introduite dans le problème de factorisation de matrice en utilisant des poids entre les individus. Les poids sont utilisés pour imposer une contrainte de régularité de l'estimateur des coefficients de métissage sur l'espace géographique. En effet, nous souhaitons que des individus proches dans l'espace géographique aient des coefficients de métissage proches. Les poids sont définis à partir des coordonnées géographiques des individus que l'on note x_i pour chaque individu i. Nous attribuons aux individus proches dans l'espace un poids plus grand que pour des individus éloignés. Les poids sont calculés en construisant un graphe complet pondéré entre les individus. Entre chaque individu i et j, nous construisons la matrice des poids du graphe $\mathbf{W}$ de la manière suivante

$$\mathbf{W}_{i,j} = \exp(-\mathrm{dist}(x_i, x_j)^2/\sigma^2), \tag{2.4}$$

où la fonction $\text{dist}(x_i, x_j)$ définit une distance entre les coordonnées géographiques x_i et x_j des individus d'indice i et j.

Ensuite, nous introduisons la matrice laplacienne associée à la matrice des poids géographiques $\mathbf{W}$. La matrice laplacienne est définie de la manière suivante

$$\mathbf{\Gamma} = \mathbf{D} - \mathbf{W}, \tag{2.5}$$

où $\mathbf{D}$ est la matrice diagonale tel que

$$\{\mathbf{D}_{i,i}\}_{i=1..n} = \left\{ \sum_{j=1}^{n} \mathbf{W}_{i,j} \right\}_{i=1..n}. \tag{2.6}$$

Par le calcul, D. Cai et al. (2011) ont montré que

$$\text{Tr}(\mathbf{Q}^T \mathbf{\Gamma} \mathbf{Q}) = \frac{1}{2} \sum_{i,j=1}^{n} \mathbf{W}_{i,j} \|\mathbf{Q}_{i,.} - \mathbf{Q}_{j,.}\|^2. \tag{2.7}$$

où Tr, la trace, est la fonction qui renvoie la somme des valeurs diagonales d'une matrice carrées. Dans notre approche, nous supposons que les individus géographiquement proches ont plus de chance d'avoir des ancêtres communs que des individus éloignés. Ainsi nous utilisons le terme défini par l'équation (2.7) pour régulariser l'estimateur de la matrice des coefficients de métissage $\mathbf{Q}$.

2.3.3 Problèmes d'optimisation des moindres carrés

L'estimation des matrices $\mathbf{Q}$ et $\mathbf{G}$ à partir de la matrice de génotype $\mathbf{Y}$ est réalisée en optimisant la fonction suivante

$$\mathcal{L}(\mathbf{Q}, \mathbf{G}) = \|\mathbf{Y} - \mathbf{Q}\mathbf{G}^T\|_{\text{F}}^2 + \alpha \text{Tr}(\mathbf{Q}^T \mathbf{\Gamma} \mathbf{Q}), \tag{2.8}$$

où la matrice $\mathbf{Q}$ appartient à Δ_Q, l'ensemble définie par les contraintes (2.1), et la matrice $\mathbf{G}$ appartient à Δ_G, l'ensemble définie par les contraintes (2.2). La notation $\|\mathbf{M}\|_{\text{F}}$ désigne la norme de Frobenius de la matrice $\mathbf{M}$. Le paramètre de régularisation α contrôle la régularité des estimations des coefficients de métissage dans l'espace géographique. Les grandes valeurs de α impliquent que les coefficients de métissage aient des valeurs proches pour les individus géographiquement proches.

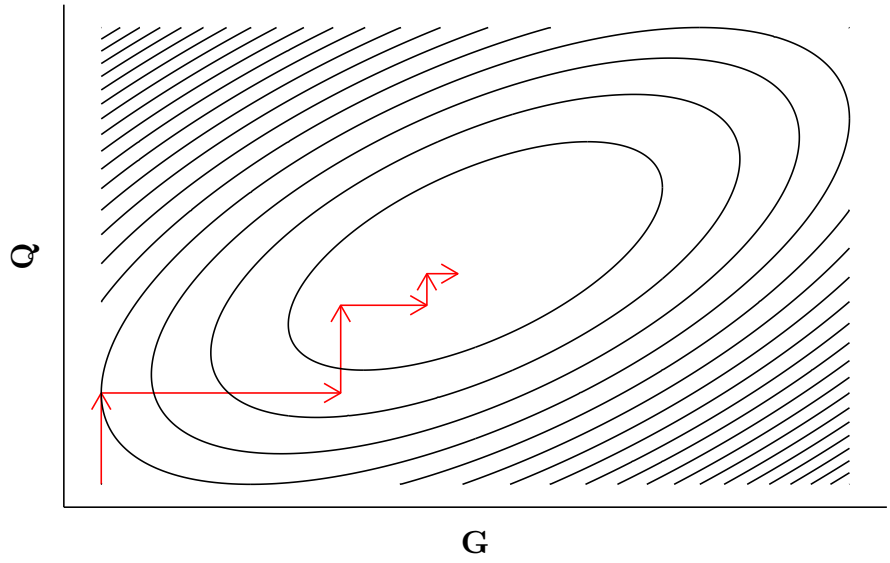

FIGURE 2.1 – **Illustration de l'algorithme de descente par blocs de coordonnées.**

2.3.4 Algorithme d'optimisation quadratique alternée (AQP)

Nous remarquons que les polyèdres Δ_Q et Δ_G sont des ensembles convexes et que la fonction $\mathcal{L}$ définie par l'équation (2.8), est convexe par rapport à chaque variable $\mathbf{Q}$ ou $\mathbf{G}$ lorsque l'autre est fixée. Nous pouvons ainsi appliquer l'algorithme de descente par blocs de coordonnées au problème afin de trouver un minimum local de la fonction $\mathcal{L}$. L'algorithme de descente par blocs de coordonnées consiste à alterner des étapes d'optimisation selon chacune des coordonnées de la fonction à optimiser (Figure 2.1). Cet algorithme converge vers un minimum local quand la fonction objectif est convexe et définie sur un ensemble convexe (BERTSEKAS, 1997). Le problème d'optimisation selon $\mathbf{G}$, quand $\mathbf{Q}$ est fixé, est un problème d'optimisation quadratique. Il en va de même quand on échange les rôles de $\mathbf{G}$ et $\mathbf{Q}$, c'est pour cela que l'algorithme est dit d'optimisation quadratique alternée (AQP).

L'algorithme APQ commence à partir de valeurs initiales pour les matrices $\mathbf{G}$ et $\mathbf{Q}$, et alterne deux étapes d'optimisation. La première étape calcule la matrice $\mathbf{G}$ tandis que la matrice $\mathbf{Q}$ est fixée. Nous supposons que $\mathbf{Q}$ est fixée et écrivons $\mathbf{G}$ sous une forme vectorielle comme ceci

$$g = \text{vec}(\mathbf{G}) \in \mathbb{R}^{K(d+1)p}.$$

La première étape de l'algorithme résout le problème d'optimisation quadratique

suivant

$$\min_{g \in \Delta_G} (-2v_Q^T g + g^T \mathbf{D}_Q g), \tag{2.9}$$

où $\mathbf{D}_Q = \mathbf{Id}_{(d+1)p} \otimes \mathbf{Q}^T\mathbf{Q}$ et $v_Q = \text{vec}(\mathbf{Q}^T\mathbf{Y})$. Ici, $\otimes$ désigne le produit Kronecker et $\mathbf{Id}_d$ est la matrice identité de taille d. La structure en blocs de la matrice $\mathbf{D}_Q$ nous permet de décomposer le problème (2.9) en p problèmes de programmation quadratiques indépendants à $K(d+1)$ variables.

Pour la deuxième étape de l'algorithme, nous considérons que $\mathbf{G}$ est la valeur obtenue après la première étape de l'algorithme, et écrivons $\mathbf{Q}$ sous une forme vectorielle

$$q = \text{vec}(\mathbf{Q}) \in \mathbb{R}^{nK} \tag{2.10}$$

La deuxième étape résout le problème de programmation quadratique suivant

$$\min_{q \in \Delta_Q} (-2v_G^T q + q^T \mathbf{D}_G q), \tag{2.11}$$

où $\mathbf{D}_G = \mathbf{Id}_n \otimes \mathbf{G}^T\mathbf{G} + \alpha\mathbf{\Gamma} \otimes \mathbf{Id}_K$ et $v_G = \text{vec}(\mathbf{G}^T\mathbf{Y}^T)$. Contrairement au problème (2.9) de la première étape, le problème (2.11) ne peut pas être séparé en plus petits problèmes. Ainsi, la deuxième étape de l'algorithme AQP nécessite de résoudre un problème de programmation quadratique à nK variables ; cela peut être très long pour les jeux de données avec beaucoup d'individus. Nous alternons ces deux étapes jusque convergence de l'algorithme AQP en un minimum local de $\mathcal{L}$.

Nous pouvons énoncer le résultat de convergence suivant.

Théoreme 1. *L'algorithme AQP qui alterne les étapes d'optimisation des problèmes* (2.9) *et* (2.11) *converge vers un minimum local de la fonction* $\mathcal{L}$ *définie par l'équation* (2.8).

Démonstration. La fonction $\mathcal{L}$ définie par l'équation (2.8) est convexe par rapport à $\mathbf{Q}$ quand $\mathbf{G}$ est fixé et inversement. De plus les ensembles de définition Δ_Q et Δ_G sont convexes. Donc, d'après le corollaire 2 établi par Grippo et Sciandrone (2000), tout point limite de l'algorithme AQP converge vers un point de minimum local de la fonction $\mathcal{L}$. □

2.3.5 Algorithme des moindres carrés alternés projetés (APLS)

Dans cette partie nous présentons l'algorithme APLS de calcul d'un minimum local de la fonction $\mathcal{L}$ définie par l'équation (2.8). Contrairement à AQP, il n'y a pas

de résultat qui garantisse la convergence de d'APLS vers un minimum local de la fonction $\mathcal{L}$. Cependant, l'algorithme APLS a une complexité algorithmique plus faible que l'algorithme AQP. L'algorithme APLS commence par initialiser au hasard les matrices $\mathbf{Q}$ et $\mathbf{G}$ puis alterne deux étapes. La matrice $\mathbf{Q}$ est calculée pendant que la matrice $\mathbf{G}$ est fixé et vice versa.

La première étape de calcul de $\mathbf{G}$ consiste à calculer

$$\mathbf{G}^\star = \arg\min \|\mathbf{Y} - \mathbf{Q}\mathbf{G}^T\|_{\mathrm{F}}^2 . \tag{2.12}$$

Cette étape peut être séparée en $(d+1)p$ (le nombre de colonnes de $\mathbf{Y}$) problèmes indépendants. De plus, cette opération peut être parallélisée. Ensuite nous projetons $\mathbf{G}^\star$ sur le polyèdre Δ_G.

Pour la seconde étape de calcul de la matrice $\mathbf{Q}$, nous commençons par calculer la matrice des vecteurs propres de la matrice laplacienne que nous notons $\mathbf{U}$, ainsi que la matrice diagonale $\mathbf{\Delta}$ formée des valeurs propres de $\mathbf{\Gamma}$. Comme la matrice laplacienne est symétrique et positive ses valeurs propres sont des nombres réels non-négatifs. D'après le théorème spectral nous avons

$$\mathbf{\Gamma} = \mathbf{U}^T\mathbf{\Delta}\mathbf{U}. \tag{2.13}$$

Après cette opération nous projetons la matrice des données $\mathbf{Y}$ sur la base des vecteurs propres de la façon suivante

$$\mathcal{P}(\mathbf{Y}) = \mathbf{U}\mathbf{Y}, \tag{2.14}$$

et, pour chaque individu, nous calculons

$$q_i^\star = \arg\min \|\mathcal{P}(\mathbf{Y})_i - \mathbf{G}q\|_2^2 + \alpha\lambda_i\|q\|_2^2, \tag{2.15}$$

où $\mathcal{P}(\mathbf{Y})_i$ est la ligne d'indice i de la matrice des données projetées, et λ_i désigne la valeur propre d'indice i de $\mathbf{\Gamma}$. Les solutions, $q_i^\star$, sont concaténées en une matrice, $\mathbf{Q}^\star$, puis la matrice $\mathbf{Q}$ est calculée par la projection de $\mathbf{U}\mathbf{Q}^\star$ sur le polyèdre Δ_Q. La complexité de la deuxième étape de APLS croît linéairement avec n, le nombre d'individus. Alors que la propriété théorique de convergence de l'algorithme AQP est perdu pour l'algorithme APLS, nous nous attendons à ce que l'algorithme APLS fournisse de bonnes approximations de l'algorithme AQP. C'est ce que nous observons dans nos expériences numériques.

2.3.6 Choix des hyperparamètres

Le choix des hyperparamètres est un problème qui est commun à toutes les méthodes d'estimation des coefficients de métissage. La méthode que nous avons présentée dans la partie précédente nécessite le choix de trois hyperparamètres : le nombre de facteurs, K, le paramètre de régularisation, α et le paramètre d'échelle géographique, σ. Nous présentons ici des méthodes qui permettent d'aider au choix de ces paramètres.

Le paramètre d'échelle géographique σ

Les tests la corrélation entre le génotype et les coordonnées géographiques sont utilisés depuis longtemps en génétique des populations. Des approches populaires reposent sur le test de Mantel (Mantel, 1967) et sur la mesure de l'auto-corrélation spatiale (Hardy, 1999 ; Epperson et T. Li, 1996). Avant d'utiliser notre méthode spatiale d'estimation des coefficients de métissage, nous proposons de choisir des valeurs de l'échelle géographique en visualisant le variogramme spatial (Cressie, 1993). Le variogramme peut être étendu aux données génétiques de la façons suivante

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \frac{1}{L} \sum_{l=1}^{(d+1)p} |\mathbf{Y}_{i,l} - \mathbf{Y}_{j,l}|, \tag{2.16}$$

où $N(h)$ est défini comme l'ensemble des individus à une distance géographique h. Visualiser le variogramme fournit des informations sur le niveau de l'auto-corrélation spatiale dans les données génétiques et donne une estimation empirique de l'échelle géographique σ. Une autre approche consiste à prendre pour paramètre d'echelle géographique la distance géographique moyenne entre les individus.

Le paramètre de régularisation α

La valeur par défaut du paramètre de régularisation α a été choisie de sorte que le terme d'attache aux données et le terme de régularisation de la fonction $\mathcal{L}$ définie par (2.8) soient du même ordre de grandeur. Ainsi, nous proposons de diviser chaque terme par sa valeur maximale. Cela revient à considérer α égal à p/λ_{max}, où λ_{max} est la plus grande valeur propre de la matrice laplacienne.

Le nombre de groupes génétiques K

Le nombre de groupes génétiques, K, peut être évalué en utilisant une technique de validation croisée fondée sur l'imputation des génotypes masqués (Wold, 1978 ; Eastment et Krzanowski, 1982 ; Alexander et Lange, 2011 ; Frichot, Mathieu et al., 2014). La procédure de validation croisée divise les entrées matricielles génotypiques en un ensemble d'apprentissage et un ensemble de test. Les probabilités de génotype pour les entrées masquées sont prédites à partir des estimations des matrices $\mathbf{G}$ et $\mathbf{Q}$, obtenues à partir d'entrées non masquées et de l'équation (2.3). Ensuite, l'erreur de prédiction est calculée en utilisant l'erreur quadratique moyenne (RMSE, Root Square Mean Error) entre le génotype prédit et le génotype réellement observé.

2.3.7 Statistique de différenciation des groupes génétiques pour détecter les locus sous adaptation locale

Les méthodes d'estimation de la structure génétique des populations arrivent à détecter des groupes génétiques grâce aux pressions évolutives (décrites dans la partie 1.1) qui provoquent la différenciation des distributions alléliques entre les différents groupes génétiques. À l'origine de cette différenciation il y a la migration, la dérive génétique et l'adaptation à l'environnement. Les locus qui ne sont pas impliqués dans un processus d'adaptation à l'environnement sont dits neutres. Nous faisons l'hypothèse qu'une large majorité des locus observés sont neutres. La migration et la dérive génétique influencent de la même façon les distributions alléliques de tous les locus neutres, induisant ainsi une différenciation typique entre les différents groupes génétiques identifiés par les méthodes d'estimation de la structure. Les locus impliqués dans l'adaptation locale peuvent être identifiés en cherchant les locus pour lesquels on observe une différenciation anormale entre les groupes génétiques (Lewontin et Krakauer, 1973). Nous proposons de calculer une statistique de différenciation entre les fréquences génomiques dans les groupes génétiques inférées par nos algorithmes pour détecter les locus sous adaptation à l'environnement.

En supposant qu'il y à K groupes génétiques, les matrices $\mathbf{Q}$ et $\mathbf{G}$ obtenues à partir des algorithmes AQP et APLS sont utilisées pour calculer la statistique de différenciation entre les groupes génétiques pour chaque locus de la façon suivante

(Martins et al., 2016)

$$F_{\mathrm{ST}}^{Q} = 1 - \sum_{k=1}^{K} q_k \frac{f_k(1-f_k)}{f(1-f)},$$

où q_k est la mesure du coefficient de métissage dans le groupe k moyenné sur tous les individus

$$q_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Q}_{i,k},$$

f_k est la fréquence d'allèle dans le groupe k au locus considéré

$$f_k = \frac{1}{d} \sum_{j=1}^{d} j\mathbf{G}_{(d+1)\ell+j,k},$$

et

$$f = \sum_{k=1}^{K} q_k f_k.$$

À un locus donné, la formule de F_{ST}^{Q} correspond à la proportion de la variance génétique qui peut être expliquée par la structure de population latente

$$F_{\mathrm{ST}}^{Q} = \frac{\sigma_T^2 - \sigma_S^2}{\sigma_T^2},$$

où σ_T^2 est la variance totale et σ_S^2 est la variance de l'erreur (Weir, 1996). En suivant la théorie ANOVA nous utilisons les statistiques $ F^Q_{ST}$ pour effectuer des tests statistiques de neutralité à chaque locus, en comparant les valeurs observées à la valeur de différenciation génomique de fond.

Le test porte sur la statistique du z-score au carré

$$z^2 = \frac{(nK)F_{\mathrm{ST}}^{Q}}{(1 - F_{\mathrm{ST}}^{Q})},$$

pour lequel une distribution du χ^2 à $K-1$ degrés de liberté est attendue sous l'hypothèse nulle. L'hypothèse nulle est ensuite calibrée empiriquement en mesurant le niveau différenciation correspondant à une différenciation neutre. Nous utilisons pour cela le facteur d'inflation génomique (Devlin et Roeder, 1999 ; François, Martins et al., 2016). Après une calibration du test, le contrôle du taux de fausse découverte est effectué en utilisant l'algorithme de Benjamini-Hochberg (Benjamini et Hochberg, 1995).

2.3.8 Implémentation en R

Les deux algorithmes APLS et AQP que nous avons présentés dans cette partie ont été implémentés en langage R dans un package que nous avons appelé `tess3r`. Le nom du package fait référence au logiciel `TESS` 2.3 qui permet aussi d'estimer l'ascendance génétique à partir de données génétiques et géographiques dans un modèle très proche de celui considéré ici. La différence majeure entre les deux logiciels est que `TESS` 2.3 utilise un algorithme MCMC pour estimer les matrices d'ascendance, alors que `tess3r` implémente des algorithmes de factorisation de matrices.

2.4 Expérimentations : données simulées et réelles

2.4.1 Données simulées

Simulations à partir de données réelles

Des sous-échantillons provenant du jeu de données réelles d'*Arabidopsis thaliana* (données décrites dans la section 2.4.2) ont été utilisés pour effectuer une analyse de la convergence et du temps de calcul des algorithmes AQP et APLS. Les temps d'exécution ont été évalués en utilisant un seul processeur informatique Intel Xeon 2.0 GHz.

Simulations coalescentes

Nous avons utilisé le programme informatique `ms` pour effectuer des simulations coalescentes de SNPs neutres et de SNPs sous adaptation à l'environnement (Hudson, 2002). Deux populations sources ont été créées à partir de la simulation du modèle à deux îles de Wright. Ensuite nous avons généré des génotypes en mélangeant les génotypes des deux populations simulées avec `ms` grâce à des proportions de mélange qui varient continuellement selon un gradient longitudinal (Durand et al., 2009 ; François et Durand, 2010) (Figure 2.2). Dans ce scénario, les individus à chaque extrémité de la zone géographique sont représentatifs de leur population d'origine, tandis que les individus au centre de la gamme partagent des niveaux intermédiaires d'ascendance dans les deux populations ancestrales. Pour ces simulations, la matrice

des coefficients de métissage, noté $\mathbf{Q}_0$, est entièrement décrite par la position des individus échantillonnés.

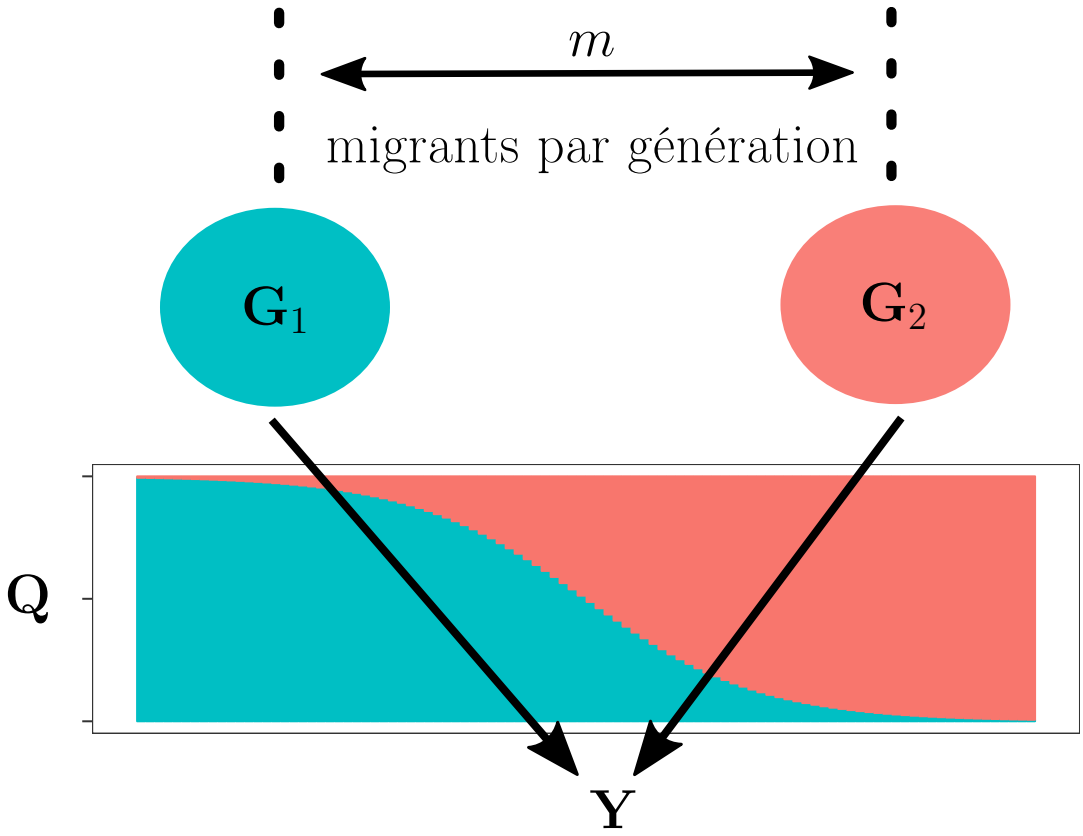

FIGURE 2.2 – **Simulation de génotypes métissés spatialement.** Les populations sources sont simulées à l'aide du programme `ms` pour obtenir les matrices de fréquence de génotypes $\mathbf{G}_1$ et $\mathbf{G}_2$. La matrice $\mathbf{Y}$ est générée en tirant des gènes des deux populations sources avec des probabilités générées selon un gradient longitudinal et stoquées dans la matrice $\mathbf{Q}$.

Les segments de chromosome neutre des populations sources ont été générés en simulant des séquences d'ADN avec une taille de population efficace de $N_0 = 10^6$ pour chaque population. Le taux de mutation par nucléotide et génération a été réglé sur $\mu = 0,25 \times 10^{-7}$; le taux de recombinaison par génération a été réglé sur $r = 0,25 \times 10^{-8}$; et le paramètre m a été choisi pour obtenir des niveaux neutres de F_{ST} compris entre des valeurs de 0.005 et 0.10. Le nombre de nucléotides pour chaque séquence d'ADN a varié entre 10k et 300k pour obtenir un nombre de locus polymorphes compris entre 1k et 200k après avoir filtré les SNPs ayant une fréquence d'allèle mineure

inférieure à 5 %. Pour créer des SNPs avec des valeurs de F_{ST} atypiques par rapport au F_{ST} des locus neutres, des segments chromosomiques ancestraux supplémentaires ont été générés en simulant des séquences d'ADN avec un taux de migration m_s inférieur à m. Les simulations ont permis de reproduire pour les SNPs sous adaptation locale les niveaux de diversité attendues lors du balayage sélectif d'un segment chromosomique particulier (Martins et al., 2016). Pour chaque simulation, la taille de l'échantillon a varié pour un nombre d'individus allant de 50 à 700.

Nous avons comparé les estimateurs des algorithmes APLS et AQP entre eux. De plus, nous avons comparé les résultats de l'algorithme APLS avec ceux obtenus par le logiciel `TESS` 2.3. Chaque programme a été exécuté 5 fois sur les mêmes données simulées en utilisant $K = 2$ groupes génétiques comme hyperparamètres. Nous avons calculé l'erreur quadratique moyenne (RMSE) entre les valeurs estimées et connues de la matrice $\mathbf{Q}$, et entre les valeurs estimées et connues de la matrice $\mathbf{G}$. Ensuite, pour évaluer le bénéfice des algorithmes spatiaux, nous avons comparé les erreurs statistiques de l'algorithme APLS aux erreurs obtenues avec la méthode sNMF qui reproduit les résultats du programme `structure` avec précision (Frichot, Mathieu et al., 2014 ; Frichot et François, 2015). Pour quantifier les performances des tests de détection de l'adaptation locale en fonction de l'intensité de la sélection environnementale, nous avons utilisé l'aire sous la courbe précision-rappel du test d'adaptation locale (AUC) pour plusieurs valeurs de l'intensité de la sélection (m/m_s).

2.4.2 Application à des écotypes européens d'*Arabidopsis thaliana*

Nous avons utilisé l'algorithme APLS pour étudier la structure de population spatiale et pour détecter les locus sous adaptation locale en considérant 214k SNPs à partir de 1 095 écotypes européens des espèces végétales *A.thaliana* (Horton et al., 2012). Le critère de validation croisée a été utilisé pour évaluer le nombre de groupes génétiques dans l'échantillon, et nous avons utilisé le variogramme spatial des données génétiques pour évaluer l'échelle de l'auto-corrélation spatiale. Nous avons utilisé les fonctions du package `tess3r` pour afficher les coefficients de métissage interpolés sur une carte géographique d'Europe. Une analyse d'enrichissement de l'ontologie des gènes utilisant le logiciel `AMIGO` (Carbon et al., 2008) a été réalisée afin d'évaluer quelles fonctions moléculaires et processus biologiques pourraient être impliqués dans

l'adaptation locale de *A.thaliana* en Europe.

2.5 Résultats

2.5.1 Analyse de la convergence et des temps d'exécution

Nous avons utilisé des simulations coalescentes de polymorphismes neutres avec un modèle spatial de mélange pour comparer les erreurs statistiques des estimations des matrices $\mathbf{Q}$ et $\mathbf{G}$ obtenues avec AQP et APLS. La référence pour la matrice $\mathbf{Q}$, noté $\mathbf{Q}_0$, est calculée à partir des proportions de mélange utilisées pour générer les génotypes à partir des deux populations sources simulées par `ms`. Pour la matrice $\mathbf{G}$, la matrice de référence, notée $\mathbf{G}_0$, est calculée à partir des fréquences empiriques des génotypes des deux populations sources. Les erreurs quadratiques moyennes (RMSE) pour les estimations des matrices $\mathbf{Q}$ et $\mathbf{G}$ diminue à mesure que la taille de l'échantillon et le nombre de locus augmentent (Figure 2.3). Pour tous les algorithmes, les erreurs statistiques sont généralement faibles lorsque le nombre de locus est supérieur à 10k. Ces résultats permettent de prouver que les deux algorithmes produisent des estimations équivalentes des matrices $\mathbf{Q}_0$ et $\mathbf{G}_0$. Les résultats permettent également de vérifier que l'algorithme APLS converge vers les mêmes estimations que celles obtenues par l'algorithme AQP, qui est garanti de converger vers un minimum de la fonction objectif $\mathcal{L}$.

Nous avons sous-échantillonné un grand jeu de données de SNPs d'écotypes de *A.thaliana* pour comparer les propriétés de convergence et les temps d'exécution des algorithmes AQP et APLS. Dans ces expériences, nous utilisons les méthodes avec 6 groupes génétiques et avons répliqué 5 fois chaque simulation. Pour un nombre d'individus allant de 500 à 600 et un nombre de 50k locus, l'algorithme APLS nécessite plus d'itérations (25 itérations) que l'algorithme AQP (20 itérations) pour converger vers sa solution (Figure 2.4). Pour un nombre de locus allant de 10k à 200k et 150 individus, nous observons des résultats similaires. Pour 50k SNPs, les temps d'exécution sont significativement plus bas pour l'algorithme APLS que pour l'algorithme AQP. Pour 50k SNPs et 600 individus, cela prend en moyenne 1,0 min pour l'algorithme APLS et 100 min pour l'algorithme AQP pour calculer leurs résultats. Pour 100k SNPs et 150 individus, il faut en moyenne 0,6 min (9,0 min) pour l'algorithme APLS (AQP) pour calculer leurs résultats. Pour le spectre de valeurs du nombre d'individus

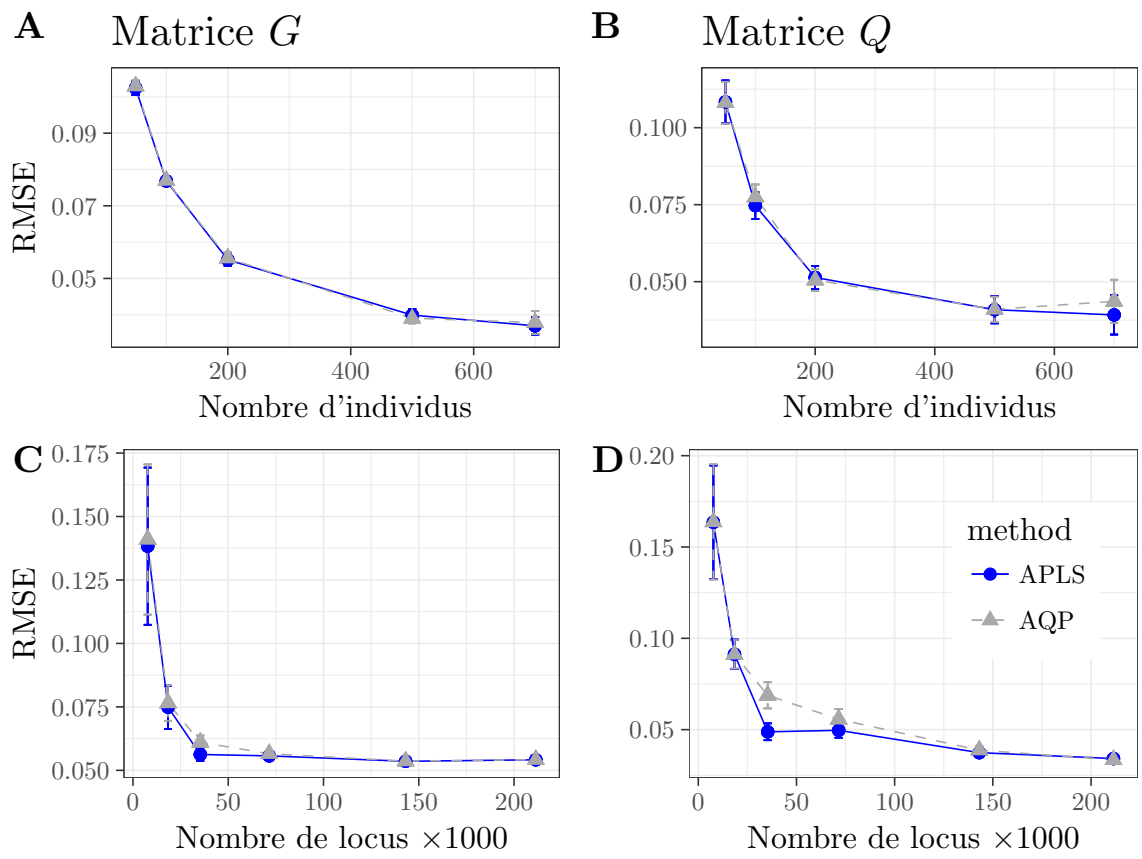

FIGURE 2.3 – **Racine carré de l'erreur quadratique moyenne (RMSE) pour les estimations des matrices Q et G**. Simulations de génotype métissé spatialement. A-B) Erreur statistique des estimations de APLS et AQP en fonction du nombre d'individu n ($p \sim 10^4$). C-D) Erreur statistique des estimations de APLS et AQP en fonction du nombre de locus p ($n = 200$)

et de locus considéré, l'implémentation de l'algorithme APLS est environ 2 à 100 fois plus rapide que celle de l'algorithme AQP.

2.5.2 Comparaison avec une méthode spatiale bayésienne : TESS

Nous avons utilisé des simulations coalescentes de polymorphismes neutres avec un modèle spatial de mélange pour évaluer les capacités de l'algorithme APLS à reproduire les estimations obtenues avec le logiciel `TESS` 2.3. Nous avons simulé deux groupes génétiques de 100 échantillons de 2000 locus pour différent niveau de différenciation moyen F_{ST}. Cela a permis de créer des jeux de données pour lesquels il est plus ou

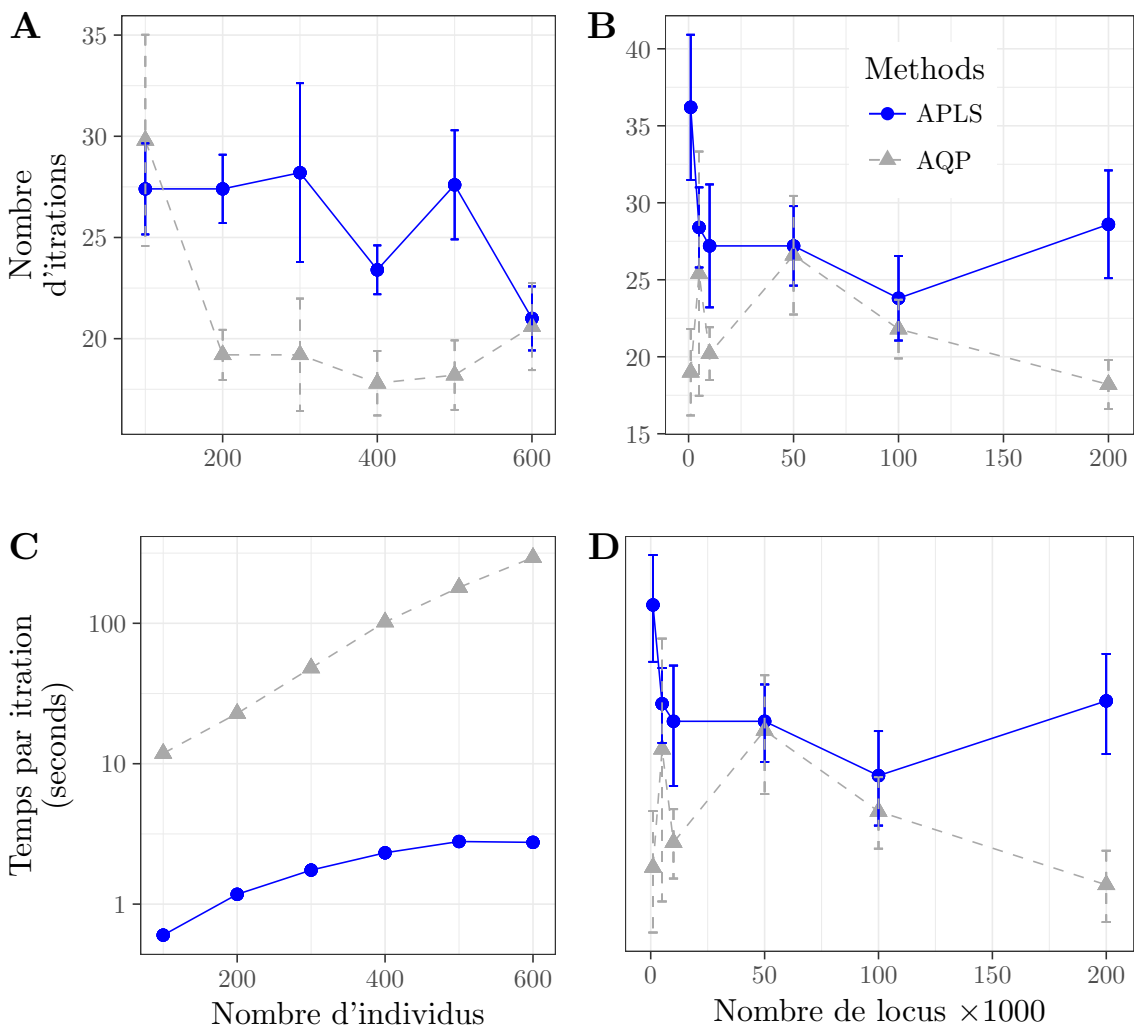

FIGURE 2.4 – **Nombre d'itérations et temps d'exécution pour les algorithmes AQP et APLS**. A-B) Nombre total d'itération avant que l'algorithmes ait atteint une solution stable. C-D) Temps de calcul d'une seul itération en secondes. Le nombre de SNPs a été fixé à $p = 50$k pour A et C. Le nombre d'individus a été fixé à $n = 150$ pour B et D.

moins difficile d'estimer la structure de population. Les erreurs statistiques, mesurées par RMSE, pour l'estimation des matrices **Q** et **G** varie entre 0.02 et 0.15 (Figure 2.5). Les erreurs statistiques augmentent à mesure que les niveaux de différenciation entre les deux populations sources diminuent, mais elles reste dans une gamme acceptable pour les valeurs de $F_{ST} > 0.016$. Dans l'ensemble, les performances statistiques sont du même ordre pour `TESS` 2.3 et APLS.

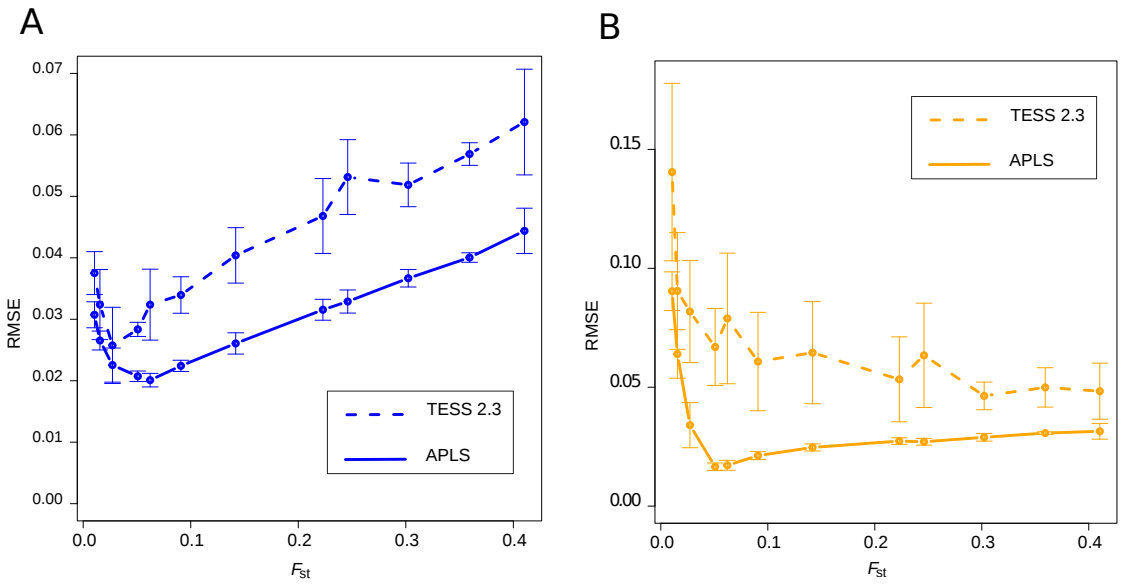

FIGURE 2.5 – **Racine de l'erreur quadratique moyenne (RMSE) pour l'estimation de Q (figure A) et G (figure B).** Simulations de génotypes métissés spatialement pour plusieurs niveaux de différenciation (F_{ST}) entre les populations sources. Les populations sources sont simulées par un modèle de Wright à deux îles et la statistique de différenciation est définie comme $1/(1+4N_0m)$ où m est le taux de migration et N_0 la taille effective de la population. Les erreurs statistiques de *TESS* 2.3 et APLS sont représentées comme des fonctions de F_{ST}.

2.5.3 Comparaison avec une méthode non spatiale : sNMF

En utilisant des simulations coalescentes de polymorphismes neutres avec un modèle spatial de mélange, nous avons comparé les estimations statistiques obtenues à partir d'un algorithme spatial (APLS) et d'un algorithme non spatial (sNMF, FRICHOT, MATHIEU et al., 2014). Pour différents niveaux de différenciation de la population ancestrale, les estimations obtenues à partir de l'algorithme spatial sont plus précises que celles obtenues en utilisant des approches non spatiales (Figure 2.6). Pour les données simulées plus grandes, la méthode spatiale arrive à détecter une structure de population plus fine que l'algorithme non spatial (Figure 2.6).

Sur les simulations avec des locus sous adaptation locale, nous avons utilisé l'aire sous la courbe de précision-rappel (AUC) pour quantifier les performances des tests de détection des locus sélectionnés reposant sur les estimations des matrices d'ascendance génétique, **Q** et **G**. De plus, nous avons calculé les AUC pour des tests de neutralité basés sur le calcul de la F_{ST} à partir des génotypes des populations sources avant

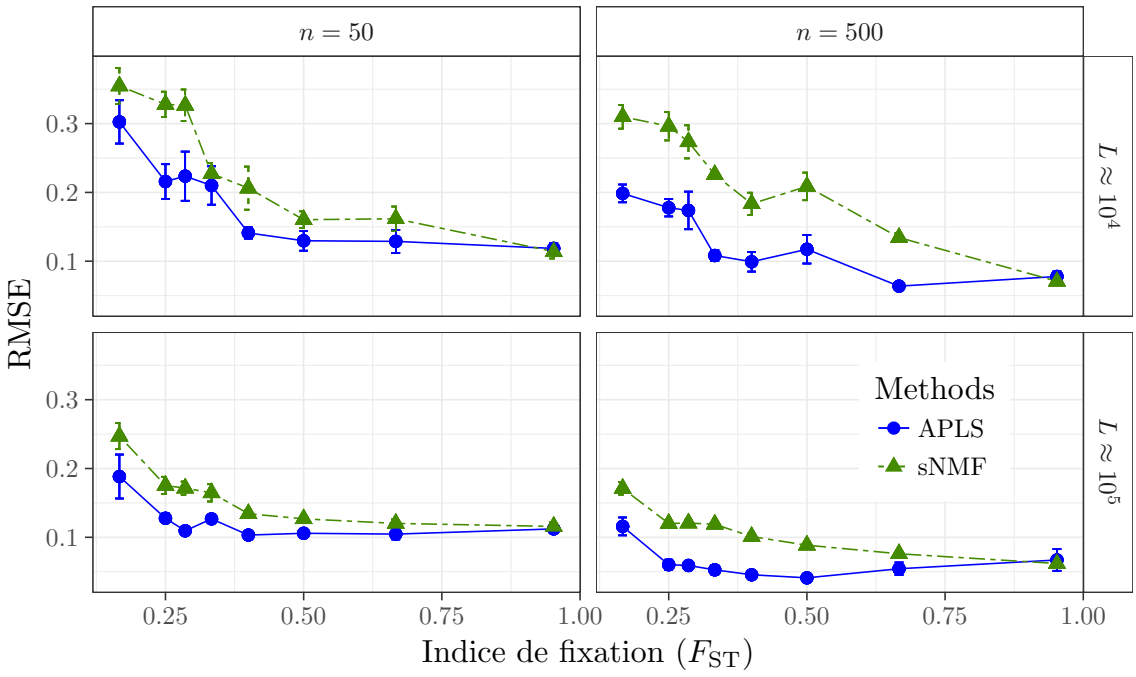

FIGURE 2.6 – **Racine de l'erreur quadratique moyenne (RMSE) pour l'estimation de Q.** Simulations de génotypes métissés spatialement pour plusieurs niveaux de différenciation (F_{ST}) entre les populations sources. Les populations sources sont simulées par un modèle de Wright à deux îles et la statistique de différenciation est définie comme $1/(1+4N_0m)$ où m est le taux de migration et N_0 la taille effective de la population. Les erreurs statistiques de sNMF et APLS sont représentées comme des fonctions de F_{ST}.

le mélange. Comme ils représentent les valeurs maximales atteignables, les AUC basés sur les génotypes des populations sources sont toujours plus élevés que ceux obtenus pour des tests basés sur des estimations des matrices d'ascendance. Pour toutes les valeurs de l'intensité de sélection (le rapport des taux de migration aux locus neutres et adaptatifs), les AUC sont plus élevées pour les méthodes spatiales que pour les méthodes non spatiales (Figure 2.7). Pour les intensités de sélection élevées, les performances des tests fondés sur les estimations des matrices d'ascendance sont proches des valeurs optimales atteintes par des tests basés sur les fréquences dans les populations sources simulées par `ms`. Ces résultats fournissent des preuves que l'inclusion de l'information spatiale dans les algorithmes d'estimation de l'ascendance améliore la détection des signatures de balayage sélectif survenant dans des groupes génétiques inconnus.

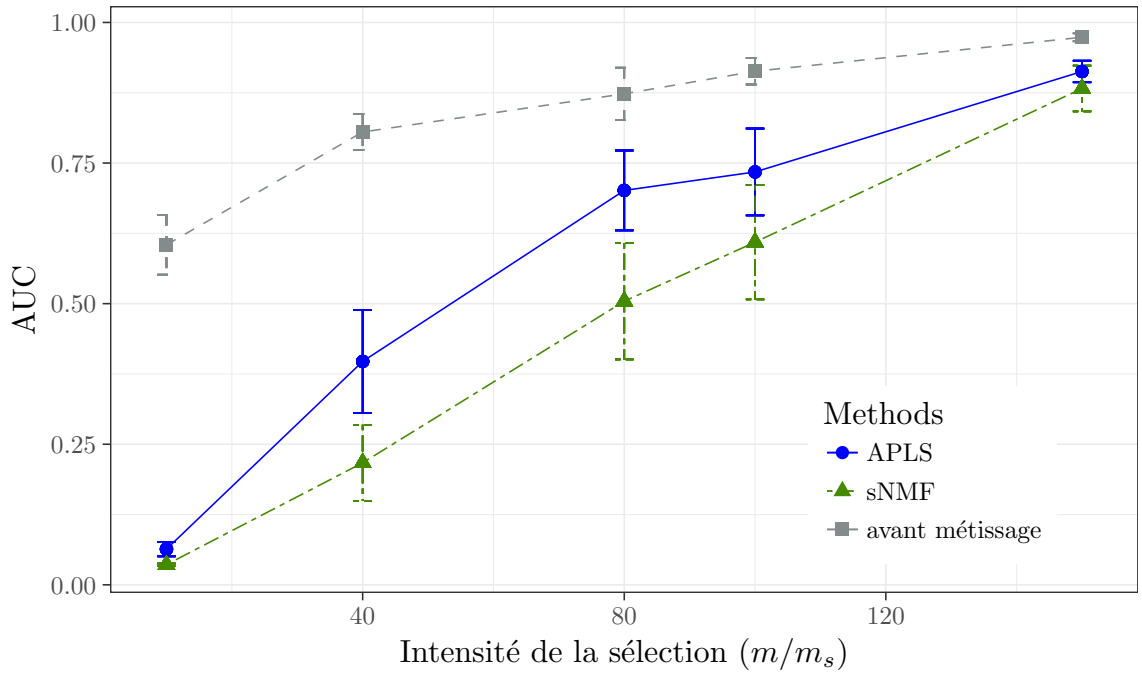

FIGURE 2.7 – **Aire sous la courbe de précision-rappel (AUC).** Tests de neutralité appliqués aux simulations de populations spatialement métissées. Nous avons calculé l'AUC pour les tests basés sur la statistique F_{ST} calculée à partir des vrais populations sources avant le métissage, à partir des estimations d'ascendance spatiale calculées avec l'algorithme APLS, l'estimations d'ascendance non spatiales (`structure`) calculées avec l'algorithme `snmf`. L'intensité de la sélection dans les populations sources, définies comme le rapport m/m_s, varie dans la fourchette de $1 - 160$.

2.5.4 Sensibilité des estimateurs aux erreurs dans les mesures spatiales

Ensuite, nous avons utilisé les jeux de données simulées afin d'évaluer la robustesse des estimations APLS pour des mesures inexactes des coordonnées spatiales. À cette fin, un bruit gaussien a été ajouté aux coordonnées géographiques vraiment observées en considérant les valeurs du rapport signal sur bruit allant de 0 à 3. Nous avons calculé le variogramme spatial pour toutes les simulations et avons constaté que la corrélation entre les données génétiques et spatiales disparaît complétement pour un niveau de signal sur bruit valant 2. Pour toutes les simulations, nous avons comparé l'erreur des estimations de la matrice $\mathbf{Q}$ fournies par APLS, relativement à celle obtenue par la méthode non spatiale sNMF. Pour de faibles niveaux d'incertitude dans les coordonnées spatiales, les erreurs statistiques des estimations APLS sont inférieures à celles de sNMF

(Figure 2.8). Pour les simulations avec $n = 500$ individus et $p = 10^5$ locus, un rapport de signal sur bruit important augmente les erreurs statistiques dans les estimations de la matrice $\mathbf{Q}$ de l'algorithme APLS. Pour un rapport signal sur bruit plus petit, les RMSE sont généralement inférieurs pour l'algorithme APLS que pour la méthode sans coordonnées spatiales. Pour les simulations avec $n = 50$ individus et $p = 10^4$ locus, les estimations APLS étaient plus précises que les estimations non spatiales. Ce résultat inattendu s'explique par des différences algorithmiques subtiles dans les programmes testés. Dans une large mesure, les estimations de l'algorithme APLS sont robustes à l'incertitude dans les mesures spatiales. Des tests graphiques standards tels qu'une analyse de variogramme peuvent aider à décider si notre algorithme spatialement explicite est utile ou non.

2.5.5 Application à des données Arabidopsis Thaliana

Nous avons utilisé l'algorithme APLS pour étudier la structure génétique de population spatiale et effectuer un balayage du génome pour les allèles adaptatifs dans des écotypes européens de l'espèce végétale *A. thaliana*. Le critère de validation croisée diminue rapidement pour $K = 1$ à $K = 3$ groupes, indiquant qu'il y a trois groupes génétiques principaux en Europe, correspondant aux régions géographiques d'Europe occidentale, d'Europe centrale et orientale et de Scandinavie septentrionale. Pour un nombre de groupes K supérieur à 4, les valeurs du critère de validation croisée diminue de manière plus lente ; ce qui indique qu'une sous-structure subtile, résultant de processus historiques complexes d'isolement par distance, pourrait également être détectée (Figure 2.9). Le variogramme spatial donne une echelle spatiale approximative de $\sigma = 150$ km (Figure 2.9). La figure 2.10 affiche l'estimation de la matrice $\mathbf{Q}$ interpolée sur une carte géographique d'Europe pour 6 groupes génétiques. L'estimation des coefficients de métissage fournit une preuve claire du regroupement des écotypes dans des groupes génétiques spatialement homogènes.

Les tests basés sur la statistique F_{ST} ont été appliqués à l'ensemble de 214k SNP pour détecter les locus sous sélection naturelle dans le génome de l'espèce *A.thaliana*. *A. thaliana* se trouve dans une large variété d'habitats, et l'adaptation locale à l'environnement est reconnue comme étant une pression importante qui façonne la diversité génétique dans l'espace (Hancock et al., 2011 ; Fournier-Level et al.,

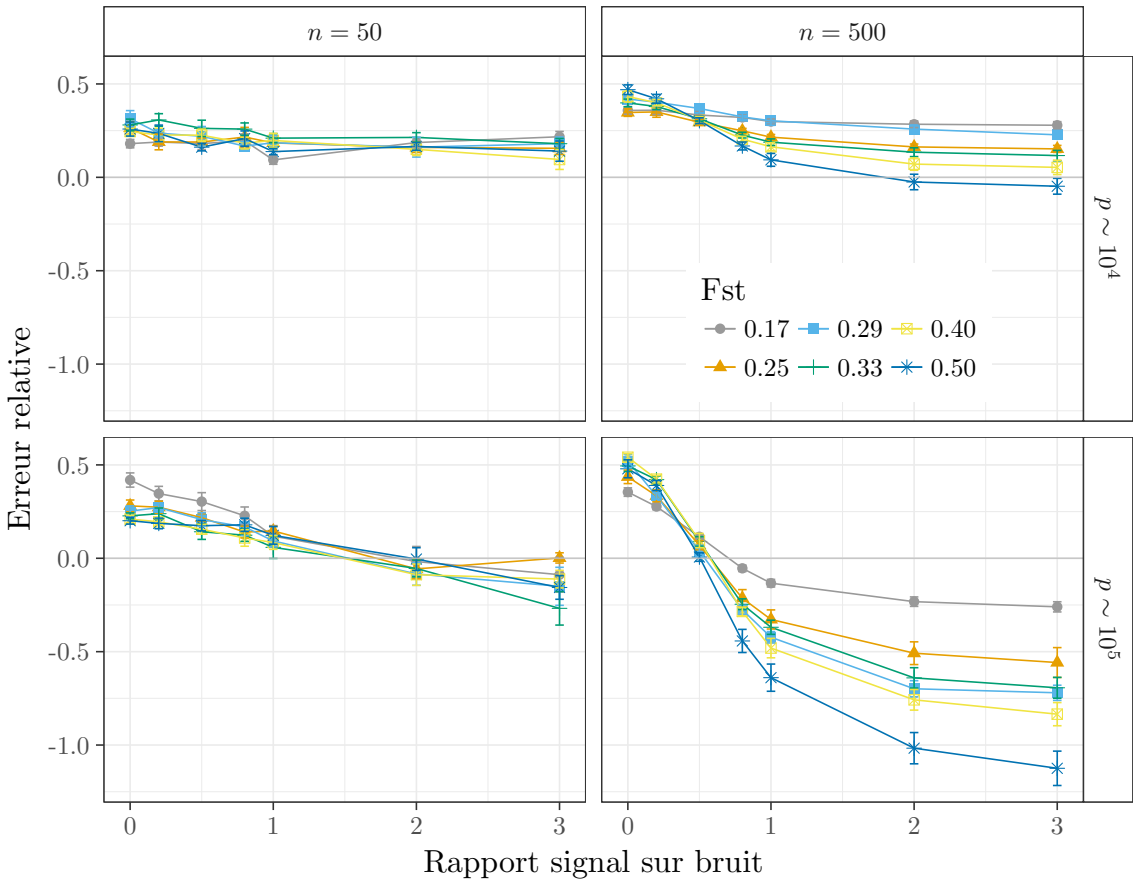

FIGURE 2.8 – **Impact de l'incertitude dans les coordonnées géographiques sur les estimations des coefficients de métissage.** Erreur statistique relative des estimations des coefficients de métissage obtenues à partir de l'algorithme APLS pour plusieurs niveaux du rapport bruit-signal et des valeurs de l'indice de fixation (F_{ST}). La méthode sNMF a été considéré comme la référence pour la méthode non spatiale (valeur 0)

2011). L'algorithme APLS a été exécuté sur les 1095 écotypes européennes de *A. thaliana* avec 6 groupes génétiques et 1.5 pour le paramètre d'échelle spatiale. En utilisant l'algorithme Benjamini-Hochberg pour contrôler le FDR à 1%, le programme a produit une liste de 12 701 SNPs candidats. Les 100 meilleurs candidats incluent des SNPs dans les gènes liés à la floraison SHORT VEGETATIVE PHASE (SVP), COP1-interacting protein 4.1 (CIP4.1) et FRIGIDA (FRI) (p-valeur $< 10^{-300}$). Ces gènes ont été détectés par des analyses antérieures de la sélection sur cet ensemble de données (HORTON et al., 2012). Nous avons réalisé une analyse d'enrichissement en ontologie des gènes en utilisant AmiGO afin d'évaluer quelles fonctions biologiques

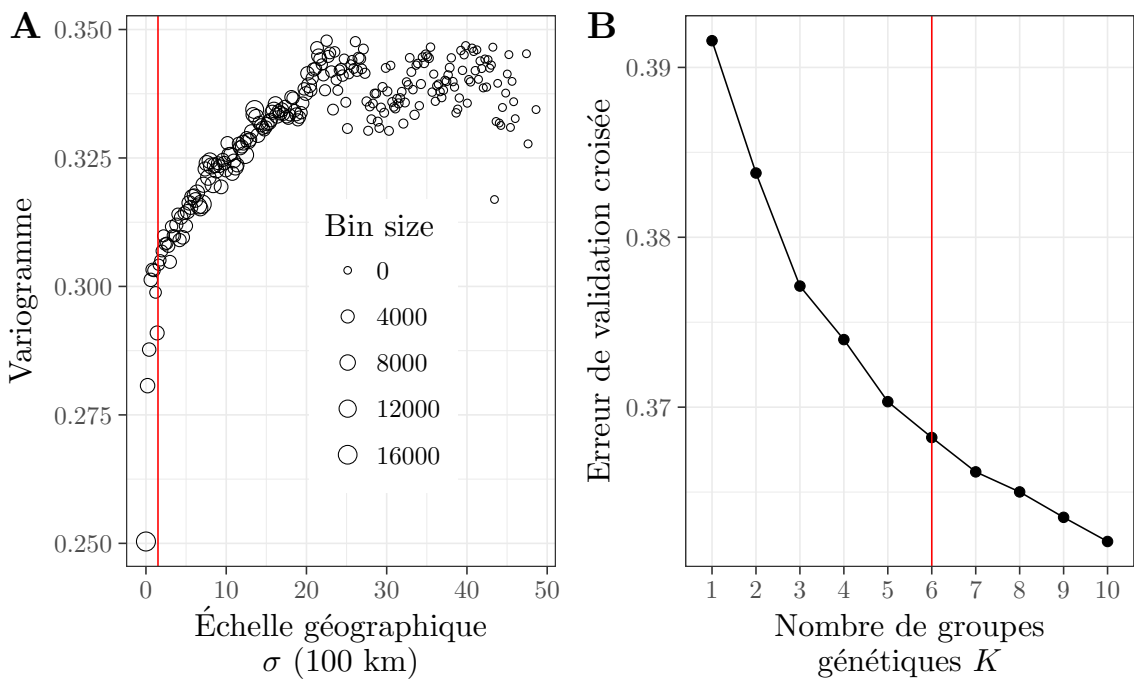

FIGURE 2.9 – **Choix de σ et K pour l'algorithme APLS**. A) Variogramme empirique pour les données *A. thaliana*. La ligne verticale rouge montre la valeur de l'échelle géographique choisie, $\sigma = 1.5$. B) Erreur de validation croisée en fonction du nombre de groupes génétiques, K. La ligne verticale rouge montre le nombre de groupes génétiques choisi, $K = 6$.

pourraient être impliquées dans l'adaptation locale en Europe. Nous avons trouvé une sur-représentation significative des gènes impliqués dans les processus cellulaires (1.06 fois plus que l'attendu, et une p-valeur égale à 0.0215 après correction de Bonferonni).

2.6 Discussion

L'inclusion de l'information géographique dans l'inférence des relations ancestrales entre les organismes est un objectif majeur des études en génétique des populations (MALÉCOT, 1948 ; EPPERSON et T. LI, 1996 ; CAVALLI et al., 1994). En supposant que des individus géographiquement proches sont plus susceptibles de partager la même ascendance que des individus dans des sites éloignés, nous avons introduit deux nouveaux algorithmes pour estimer les proportions d'ascendance à l'aide d'informations géographiques et génétiques. Sur la base des problèmes de moindres carrés, les nouveaux algorithmes combinent des approches de factorisation matricielle et de statistiques

spatiales pour fournir des estimations précises des coefficients de métissage individuels et des fréquences d'allèle dans les groupes génétiques. Les deux algorithmes partagent de nombreuses similitudes, mais ils diffèrent dans les approximations qu'ils font pour diminuer la complexité algorithmique. Plus précisément, l'algorithme AQP alterne des résolutions de problèmes d'optimisation quadratique alors que l'algorithme APLS lève les contraintes des problèmes d'optimisation afin de les transformer en problèmes des moindres carrés. La complexité algorithmique de l'algorithme APLS augmente linéairement avec le nombre d'individus dans l'échantillon en ayant la même précision statistique que l'algorithme AQP plus lent.

Pour mesurer le bénéfice de l'utilisation d'algorithmes spatiaux, nous avons comparé les erreurs statistiques observées pour les algorithmes spatiaux avec celles observées pour les algorithmes non spatiaux. Dans nos expériences numériques les erreurs des méthodes spatiales sont inférieures à celles observées avec des méthodes non spatiales, et les algorithmes spatiaux ont permis de détecter une structure de population plus subtile. En outre, nous avons mis en place un test de détection de la sélection reposant sur les estimations spatiales des matrices **Q** et **G** (Martins et al., 2016) ; et nous avons observé que les tests rejettent la neutralité des locus avec plus de précision que ceux fondés sur des approches non spatiales. Ainsi, l'information spatiale a contribué à améliorer la détection des signatures de balayage sélectif survenant dans des populations sources avant les événements de mélange. Nous avons appliqué les tests de neutralité afin d'effectuer un balayage du génome pour la sélection dans des écotypes européens de l'espèce végétale *A.thaliana*. Le scan du génome a confirmé la preuve de la sélection des gènes liés à la floraison *CIP4.1*, *FRI* et *DOG1* différenciant la Fenno-Scandinavie du nord-ouest de l'Europe (Horton et al., 2012).

L'estimation des coefficients de métissage, en utilisant des algorithmes rapides qui étendent des approches non spatiales telles que `structure`, a été intensément discutée au cours des dernières années (Wollstein et Lao, 2015). Dans ces améliorations, les approches spatiales ont reçu moins d'attention que les approches non spatiales. Dans cette étude, nous avons présenté un cadre conceptuelle permettant de développer des méthodes rapides d'estimation de l'ascendance spatiale. Ces méthodes rapides sont implémentées dans le package `tess3r` qui propose une pipeline intégrée pour l'estimation et la visualisation de la structure génétique de population et pour la recherche de locus responsables de l'adaptation locale. La complexité algorithmique

de nos algorithmes permet à leurs utilisateurs d'analyser des échantillons comprenant des centaines à des milliers d'individus. Par exemple, l'analyse de plus d'un millier de génotypes *A.thaliana*, chacun incluant plus de 210k SNP, n'a pris que quelques minutes à l'aide d'un seul CPU.

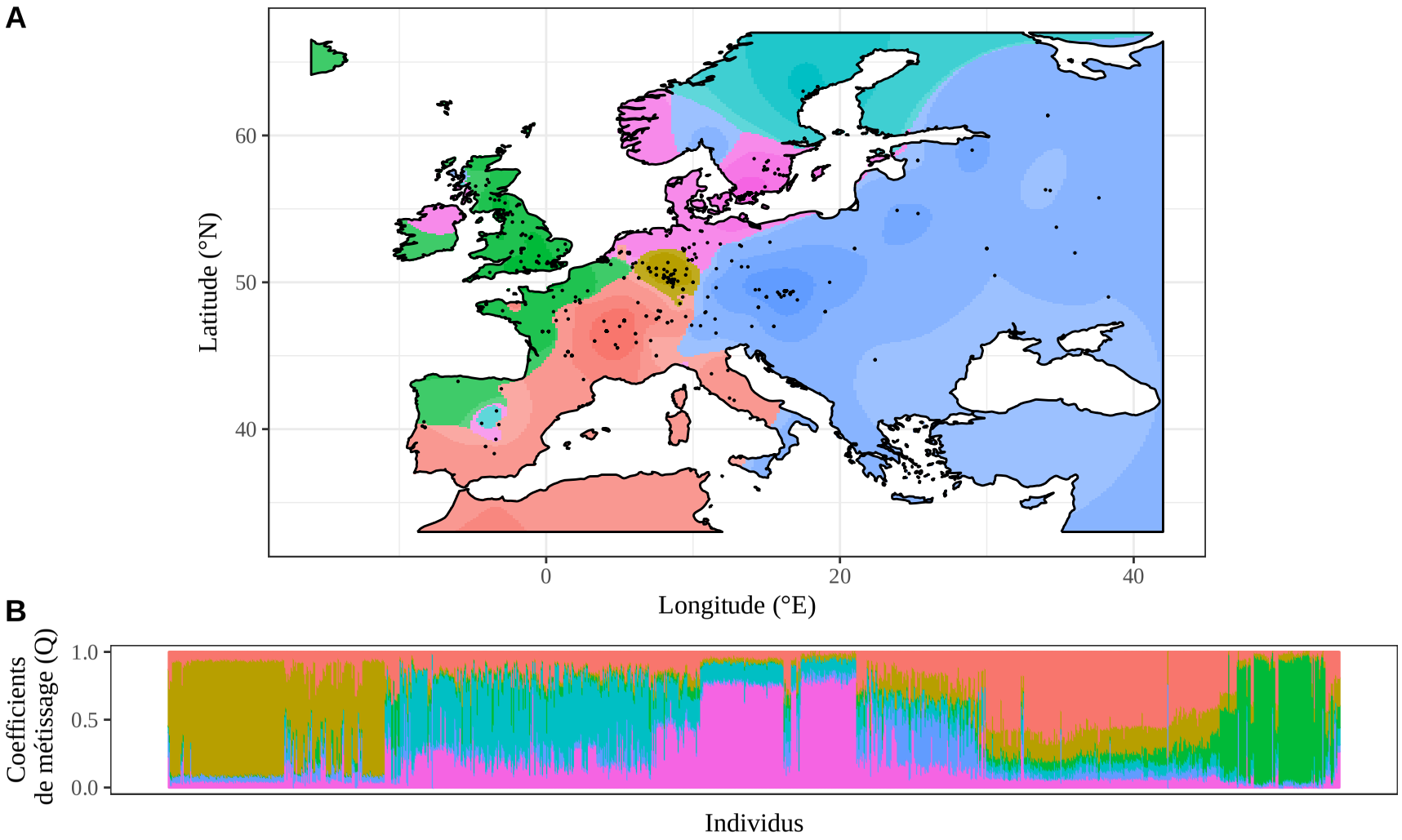

FIGURE 2.10 – *A. thaliana* **coefficients de métissage**. Estimation des coefficients de métissage calculée par l'algorithme APLS avec $K = 6$ groupes génétiques et $\sigma = 1.5$ pour le paramètre d'échelle géographique. A) Carte géographique des coefficients de métissage. B) Barplot des coefficients de métissage.

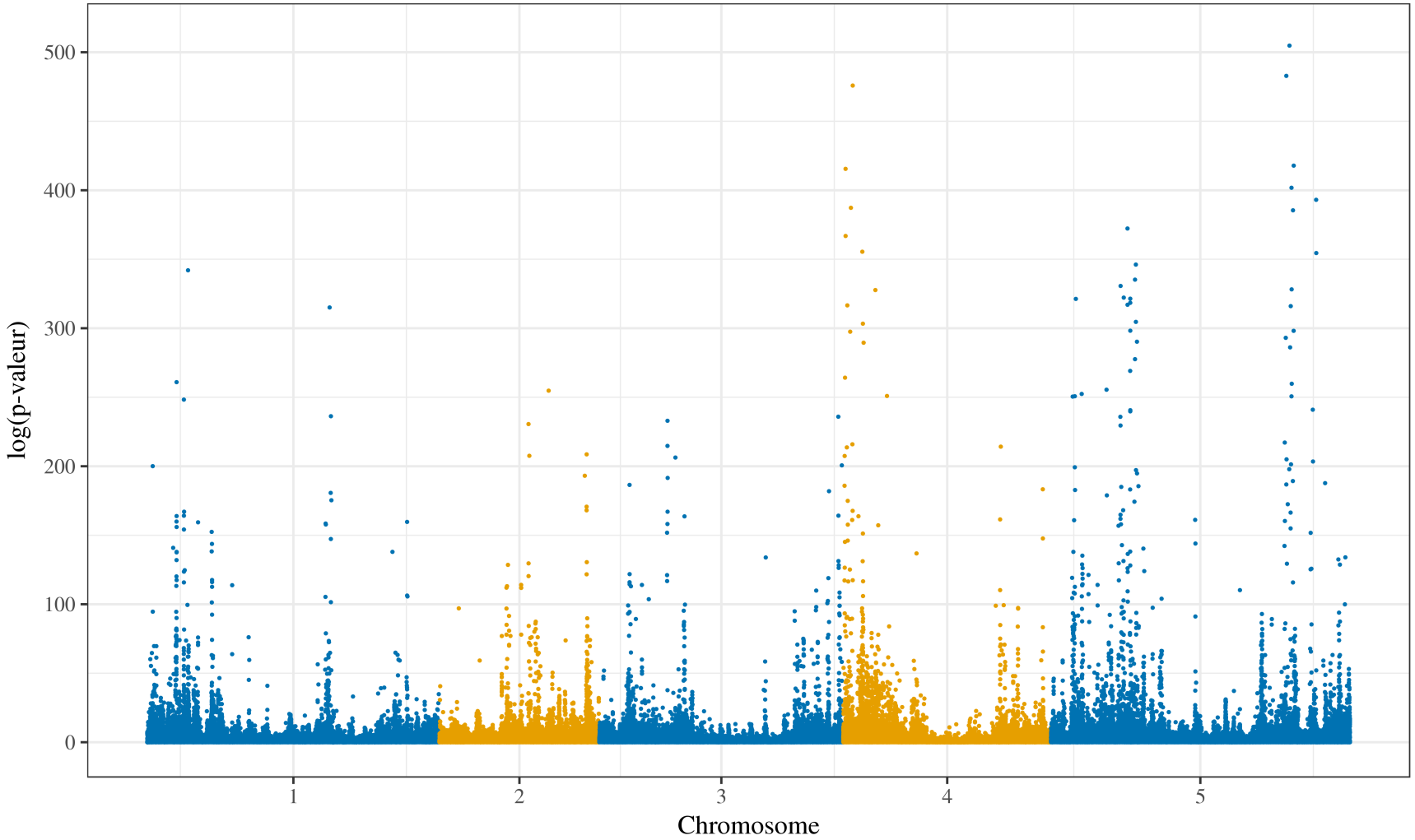

FIGURE 2.11 – **Détection de locus sous adaptation locale sur des écotypes européennes de** *A. thaliana*. Manhattan plot de $-\log(p\text{-value})$. Les p-valeurs ont été calculées à partir de la structure de la population estimée par l'algorithme APLS avec $K = 6$ groupes génétiques et $\sigma = 1,5$ pour le paramètre d'échelle géographique.

Chapitre 3

Algorithmes d'estimation pour les modèles de régression à facteurs latents

3.1 Résumé

Les études d'association sont massivement utilisées en biologie. Elles permettent par exemple de découvrir les causes génétiques d'une maladie ou bien les gènes impliqués dans un processus d'adaptation au climat. Lors de ces études, nous sommes amenés à tester les corrélations entre un grand nombre de variables, comme par exemple des SNPs, et une variable d'intérêt, comme par exemple une maladie. Dans certaines de ces études, des sources de variation indésirable, telles que la structure de populations, ou bien les conditions expérimentales, peuvent sévèrement biaiser les résultats des tests. Une méthode communément utilisée consiste à modéliser les sources de variation indésirable à l'aide de variables latentes. Dans ce chapitre, nous introduisons deux nouveaux algorithmes qui permettent d'estimer les variables latentes afin de corriger les études d'association pour les sources de variation indésirable. Nous appelons ces variables latentes les facteurs de confusion pour l'étude d'association. Nous utilisons l'approche classique qui consiste à ajouter un terme factoriel au modèle de régression des variables étudiées par des variables d'intérêts. Nos deux algorithmes reposent sur différentes régularisations portant sur les effets d'intérêt. Le premier algorithme utilise une régularisation en norme L_2, tandis que le second utilise un régularisation en norme L_1. Pour chaque algorithme, nous démontrons leur convergence vers le minimum de leur fonction objectif respective. Nous montrons sur des simulations

numériques que nos algorithmes reproduisent les performances d'autres algorithmes de correction pour les études d'association. Nous montrons enfin que nos algorithmes permettent de reproduire des résultats d'une étude d'association entre des variants génétiques humain et la maladie cœliaque, ainsi qu'une étude d'association entre des niveaux de méthylation de l'ADN et la polyarthrite rhumatoïde. Nous présentons enfin des résultats trouvés par nos algorithmes sur un exemple d'étude d'association entre des génotypes et un gradient climatique. Nos algorithmes sont implémentés dans le package R, `lfmm`.

3.2 Introduction

Au cours de la dernière décennie, des études d'association à grande échelle ont été utilisées pour identifier des gènes candidats associés à une maladie particulière ou un trait phénotypique d'intérêt. Selon le type des marqueurs moléculaires examinés dans les génomes ou dans les cellules, plusieurs catégories d'étude d'association ont été menées pour détecter des corrélations significatives entre les marqueurs et le phénotype. Par exemple, les études d'association à l'échelle du génome (GWAS genome-wide association studies) se concentrent sur des polymorphismes à un seul nucléotide (SNP pour single-nucleotide polymorphism) en examinant des variants génétiques chez différents individus (Balding, 2006). Les GWAS ont été étendues à des études d'association à l'échelle de l'épigénome (EWAS epigenome-wide association studies) qui mesurent les niveaux de méthylation de l'ADN chez différents individus pour des associations entre la variation épigénétique et des phénotypes (Rakyan et al., 2011). Des approches similaires ont été appliquées à la caractérisation de la variation observée dans l'ARN par rapport à différents environnements, traitements médicaux, phénotypes ou maladies (Slonim, 2002). D'autres exemples d'études d'association incluent des études d'association génétique-environnement (GEAS) dans lesquelles des locus sont testés pour leurs corrélations avec des gradients écologiques afin de détecter des signatures de sélection naturelle (Rellstab et al., 2015). En peu de temps, les études d'association ont permis des progrès considérables dans l'identification des variants génétiques, qui confèrent une susceptibilité aux maladies, ainsi qu'une compréhension plus approfondie de l'évolution des génomes en réponse à la sélection naturelle.

3.2.1 Méthodes de correction pour les facteurs latents

Dans cette section, nous nous plaçons dans le cadre méthodologique des modèles de régression linéaire. Il s'agit d'un cadre très utilisé en étude d'association que nous pouvons formaliser de la façon suivante

$$\mathbf{Y}_j = \mathbf{X}b_j + \mathbf{E}_j \tag{3.1}$$

où $\mathbf{Y}_j$ est la matrice des observations de la variable d'indice j pour n individus, typiquement constituée de variants génétiques. Le coefficient b_j représente l'effet de la variable $\mathbf{X}$ sur $\mathbf{Y}_j$. La matrice $\mathbf{E}_j$ est la matrice de l'erreur résiduelle. Il arrive parfois que l'on fasse la régression dans l'autre sens, la régression s'écrit alors

$$\mathbf{X} = \mathbf{Y}_j a_j + \mathbf{E}_j^{'}, \tag{3.2}$$

où j représente l'effet de $\mathbf{Y}_j$ sur $\mathbf{X}$. Dans la suite nous discutons de régression uniquement dans le sens de l'équation (3.1). L'objectif est de trouver les coefficients b_j qui sont significativement différents de zéro. Dans ce cas, nous disons que $\mathbf{Y}_j$ est associée à $\mathbf{X}$. Comme nous l'avons évoqué dans la partie précédente, avec cette approche il est possible qu'une ou plusieurs variables latentes soient corrélées à la fois à $\mathbf{Y}$ et à $\mathbf{X}$. Dans ce cas, si nous ne considérons pas les variables latentes comme variables explicatives de la régression, nous détectons un grand nombre de variables $\mathbf{Y}_j$ significativement corrélées à $\mathbf{X}$. Nous allons maintenant présenter différentes approches de correction des études d'association pour les facteurs de confusion.

Estimation des facteurs latents a priori

Une première approche consiste à estimer les variables latentes en faisant une analyse factorielle de $\mathbf{Y}$. On fait l'analyse factorielle a priori et sans prendre en compte la variable $\mathbf{X}$. Les variables latentes sont ensuite ajoutées au modèle de régression avec les autres variables explicatives, de sorte que nous avons

$$\mathbf{Y}_j = \mathbf{X}b_j + \bar{\mathbf{U}}\mathbf{V}_j^T + \mathbf{E}_j \tag{3.3}$$

où $\bar{\mathbf{U}}$ est la matrice des variables latentes calculée a priori et $\mathbf{V}_j$ la matrice des effets des variables latentes sur $\mathbf{Y}_j$. Par exemple, les méthodes EIGENSTRAT et Refactor calculent les variables latentes à l'aide de l'analyse en composantes principales (ACP) de la matrice $\mathbf{Y}$ (Price et al., 2006 ; Rahmani et al., 2016).

Les modèles mixtes

Une autre approche de correction pour les facteurs de confusion est le modèle mixte. Dans un tel modèle, on ajoute un effet aléatoire au modèle de la régression

$$\mathbf{Y}_j = \mathbf{X}\mathbf{B}^T + \mathbf{Z}\mathbf{\Gamma}_j + \mathbf{E}_j, \tag{3.4}$$

où $\mathbf{Z}$ est une matrice fixée à l'avance ("design matrix" en anglais) et $\mathbf{\Gamma}_j$ est la matrice des effets aléatoires à estimer. Dans les modèles mixtes, on suppose de plus que la matrice de covariance de l'effet aléatoire est connue. La matrice de covariance doit correspondre à la variance du facteur de confusion ; elle est en général calculée à partir de $\mathbf{Y}$. Les modèles mixtes ont été largement utilisés pour les GWAS (Kang et al., 2008 ; X. Zhou et Stephens, 2014). Dans les GWAS, les rôles de $\mathbf{Y}_j$ (un variant génétique) et $\mathbf{X}$ (un phénotype) sont inversés dans l'équation (3.4) ; et l'effet aléatoire permet d'expliquer la variation de $\mathbf{Y}_j$ qui est due à la structure de population. Dans ce cas, la matrice de covariance est estimée a priori à partir les données génétiques.

Les modèles mixtes à facteurs latents (LFMM latent factor mixed model)

Nous introduisons maintenant les modèles mixtes à facteurs latents. Pour de tels modèles, l'équation de régression peut s'écrire comme ceci :

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T + \mathbf{E}, \tag{3.5}$$

Dans cette équation $\mathbf{U}$ est la matrice des variables latentes et $\mathbf{V}$ est la matrice des effets des facteurs latents. La différence majeure entre LFMM et les autres modèles est que l'on ne suppose rien a priori sur les facteurs de confusion. Dans LFMM, l'estimation des variables latentes $\mathbf{U}$ et des effets de la variable explicative $\mathbf{B}$ se font en même temps. Les matrices $\mathbf{U}$ et $\mathbf{V}$ permettent d'apprendre les variations systématiques observées dans la matrice $\mathbf{Y}$ tandis que la matrice des effets $\mathbf{B}$ permet d'apprendre l'effet de $\mathbf{X}$ sur les colonnes de $\mathbf{Y}$. Il existe différentes méthodes pour estimer les paramètres de LFMM. On distingue d'abord des approches qui reposent sur des algorithmes de Monte-Carlo (Frichot, Schoville et al., 2013 ; Carvalho et al., 2008). Ces approches reposent sur une modélisation bayésienne qui permet d'échantillonner les lois a posteriori des paramètres. L'avantage des méthodes MCMC est qu'elles permettent d'estimer la variance du paramètre $\mathbf{B}$ grâce au bootstrap bayésien (Rubin et al., 1981),

et de faire un test de significativité statistique. Certaines approches reposent sur des algorithmes EM (Expectation Maximisation) (Friguet et al., 2009 ; Agarwal et B.-C. Chen, 2009 ; Y. Zhou et al., 2016). Ces approches sont souvent plus rapides que les méthodes utilisant des algorithmes de Monte-Carlo. Enfin, d'autres approches reposent sur une estimation des variables latentes à partir d'une transformation de $\mathbf{Y}$ (Gerard et Stephens, 2017b ; Wang et al., 2017 ; Leek et Storey, 2007). Cette transformation a pour but de séparer la variation de $\mathbf{Y}$, expliquée par les variables latentes, de celle expliquée par $\mathbf{X}$. Parmi les méthodes reposant sur une transformation des données, on distingue des autres les méthodes dites à contrôle négatif qui supposent connu un sous-ensemble de colonnes de $\mathbf{Y}$ qui ne sont pas associées à $\mathbf{X}$. Les méthodes à contrôle négatif utilisent les variables dites nulles pour estimer les variables latentes. L'estimation des variables latentes pour corriger les études d'association est un problème très vaste et aucune méthode ne s'est imposée comme étant la méthode référence.

3.2.2 Plan du chapitre

Nous proposons, dans ce chapitre, deux algorithmes d'estimation rapide et efficace des paramètres du modèle LFMM. Nos deux algorithmes d'estimation consistent à isoler la variation de $\mathbf{Y}$, expliquée par les variables latentes, de celle expliquée par les variables explicatives $\mathbf{X}$. Les méthodes que nous présentons sont comparables aux méthodes sva (Leek et Storey, 2007) et cate (Wang et al., 2017), qui procèdent d'une façon très similaire. Nous décrivons plus en détail les méthodes cate et sva dans la partie 3.4. Chacun des algorithmes que nous présentons découle de l'optimisation d'une fonction objectif. Nous montrons que nos algorithmes d'estimation convergent vers le point de minimum global de leur fonction objectif respective. Enfin, nous comparons nos méthodes à sva et cate sur des simulations numériques ainsi que pour des exemples de GWAS, EWAS et GEAS.

3.3 Nouvelles méthodes de correction pour les facteurs de confusion

3.3.1 Modèle mixte à facteurs latents

Dans cette partie, nous introduisons les notations du modèle mixte à facteurs latents que nous utilisons pour corriger les tests d'association :

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T + \mathbf{E}. \tag{3.6}$$

Dans cette équation, $\mathbf{Y}$ est la matrice de taille $n \times p$ qui rassemble les observations de p variables pour n individus. Par exemple, la matrice $\mathbf{Y}$ peut contenir des SNPs, des niveaux de méthylation ou bien des niveaux d'expression génique. Nous appelons la matrice $\mathbf{Y}$ la matrice des variable expliquées. La matrice $\mathbf{X}$, de taille $n \times d$, regroupe toutes les variables explicatives. Les variables explicatives ne sont pas obligatoirement toutes des variables d'intérêt pour l'étude d'association ; des variables explicatives supplémentaires peuvent être ajoutées pour améliorer le modèle. Les variables explicatives peuvent être par exemple un phénotype, comme une maladie, ou un gradient environnemental, comme la température d'un habitat. La matrice des effets de $\mathbf{X}$ sur $\mathbf{Y}$, de taille $p \times d$, est notée $\mathbf{B}$. Si l'on suppose qu'il y a K variables latentes, la matrice $\mathbf{U}$ est la matrice des K variables latentes et la matrice $\mathbf{V}$ représente les effets des variables latents sur $\mathbf{Y}$. Les matrices $\mathbf{V}$ et $\mathbf{U}$ sont respectivement la matrice des effets latents (appelée aussi matrice des axes factoriels), de taille $p \times K$, et la matrice des variables latentes, de taille $n \times K$. Enfin la matrice $\mathbf{E}$ est la matrice d'erreur résiduelle, de taille $n \times p$.

Dans un premier temps, nous remarquons que les matrices $\mathbf{U}$ et $\mathbf{V}$ ne sont pas définies de façon unique. En effet, comme ces deux matrices sont multipliées entre elles dans l'équation (3.6), les matrices $\mathbf{U}$ et $\mathbf{V}$ sont définies à une matrice inversible près

$$\mathbf{U}\mathbf{V}^T = \mathbf{U}\mathbf{R}\mathbf{R}^{-1}\mathbf{V}^T \tag{3.7}$$

où $\mathbf{R}$ est une matrice inversible de taille $K \times K$. Nous considérons donc la matrice

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T \tag{3.8}$$

et nous appelons la matrice $\mathbf{W}$ la matrice latente. Si l'on suppose qu'il y a K variables latentes linéairement indépendantes, cela équivaut à faire l'hypothèse que la matrice latente $\mathbf{W}$ est de rang K. Dans la suite, nous considérons $\mathbf{U}$ et $\mathbf{V}$ comme étant les matrices uniques obtenues grâce à l'analyse en composantes principales de la matrice latente $\mathbf{W}$.

3.3.2 Estimateur des moindres carrés régularisé en norme L_2

Dans cette partie, nous présentons un algorithme d'estimation des paramètres du modèle défini par l'équation (3.6). L'algorithme d'estimation est fondé sur un problème des moindres carrés régularisé en norme L_2. Nous montrons que cet algorithme calcule un minimum global du problème d'optimisation des moindres carrés régularisé en norme L_2.

Fonction objectif

Afin d'estimer les paramètres $\mathbf{U}$, $\mathbf{V}$ et $\mathbf{B}$ de LFMM, nous définissons la fonction objectif de type ridge suivante

$$\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{U}\mathbf{V}^T - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \frac{\lambda}{2} \|\mathbf{B}\|_2^2 \tag{3.9}$$

où $\|.\|_F$ est la norme de Frobenius, $\|.\|_2$ est la norme L_2 et λ le paramètre de régularisation. Le premier terme de la fonction $\mathcal{L}_{\text{ridge}}$ est le terme d'attache aux données. Si il n'y a pas de variable explicative $\mathbf{X}$, le terme d'attache aux données correspond à la fonction objectif de l'analyse en composantes principales. Le deuxième terme de la fonction $\mathcal{L}_{\text{ridge}}$ est le terme de régularisation. Ce terme est indispensable pour séparer les variations de $\mathbf{Y}$ expliquées par les variables latentes de celles expliquées par les variables explicatives. En effet, si $\lambda = 0$ alors, pour toute matrice $\mathbf{P}$, de taille $d \times p$, nous avons

$$\mathcal{L}_{\text{ridge}}(\mathbf{U} - \mathbf{X}\mathbf{P}, \mathbf{V}^T, \mathbf{B} + \mathbf{V}\mathbf{P}^T) = \mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}^T, \mathbf{B}).$$

Nous voyons que les points du minimum de la fonction objectif ne sont pas définis de manière univoque pour notre problème quand le paramètre de régularisation est nul.

Algorithme de minimisation de la fonction objectif $\mathcal{L}_{\text{ridge}}$

Afin d'estimer les paramètres de LFMM minimisant la fonction $\mathcal{L}_{\text{ridge}}$, nous commençons par calculer la décomposition en valeurs singulières de $\mathbf{X}$

$$\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T,$$

où $\mathbf{Q}$ une matrice unitaire de taille $n \times n$, $\mathbf{R}$ une matrice unitaire de taille $d \times d$ et $\mathbf{\Sigma}$ une matrice de taille $n \times d$ contenant les valeurs singulières $\{\sigma_j\}_{j=1..d}$ de $\mathbf{X}$. Les

estimateurs sont calculés de la façon suivante

$$\hat{\mathbf{U}}\hat{\mathbf{V}}^T = \mathbf{Q}\mathbf{D}_\lambda^{-1}\mathrm{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{3.10}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T), \tag{3.11}$$

où $\mathrm{svd}_K(\mathbf{A})$ est la meilleure approximation de rang K de la matrice $\mathbf{A}$, donnée par la décomposition en valeurs singulières et $\mathbf{Id}_d$ est la matrice identité de taille $d \times d$. La matrice $\mathbf{D}_\lambda$ est la matrice diagonale de taille $n \times n$ qui contient les termes diagonaux suivants

$$\{\mathbf{D}_{\lambda,i,i}\}_{i=1..n} = \left\{\sqrt{\frac{\lambda}{\lambda+\sigma_1^2}}, ..., \sqrt{\frac{\lambda}{\lambda+\sigma_d^2}}, 1, ..., 1\right\}.$$

Notons que l'estimation de la matrice latente $\hat{\mathbf{U}}\hat{\mathbf{V}}^T$ dans l'équation (3.10) fait intervenir la matrice de changement de base $\mathbf{Q}$. Les d premiers axes de la base canonique transformée par $\mathbf{Q}$ forment une base orthonormale de l'espace vectoriel engendré par les variables explicatives $\mathbf{X}$. La matrice diagonale $\mathbf{D}_\lambda$ a pour effet de ramener vers zéro la composante qui appartient à l'espace engendré par $\mathbf{X}$. Si λ tend vers zéro, multiplier $\mathbf{Y}$ par $\mathbf{D}_\lambda\mathbf{Q}^T$ revient à prendre le résidu d'une régression linéaire de $\mathbf{Y}$ par $\mathbf{X}$; on enlève alors toute la part de variance expliquée par $\mathbf{X}$. À la limite, $\mathbf{D}_\lambda$ n'est plus inversible. Si λ est très grand, $\mathbf{D}_\lambda$ tend vers la matrice identité. Dans ce cas, le calcul de $\hat{\mathbf{U}}\hat{\mathbf{V}}$ revient à faire une analyse en composantes principales de la matrice des variables expliquées $\mathbf{Y}$. Nous expliquons plus en détail dans la section 3.3.5 comment choisir l'hyperparamètre λ.

Les estimateurs des paramètres régularisés en norme L_2 sont justifiés par le théorème suivant.

Théoreme 2. *Pour λ strictement supérieur à zéro, les estimateurs des paramètres de LFMM régularisés en norme L_2, définis par* (3.10) *et* (3.11)*, définissent un minimum global de la fonction objectif $\mathcal{L}_{\text{ridge}}$.*

Démonstration. On veut trouver $\hat{\mathbf{U}} \in \mathbb{R}^{n\times K}$, $\hat{\mathbf{V}} \in \mathbb{R}^{p\times K}$ et $\hat{\mathbf{B}} \in \mathbb{R}^{p\times d}$ correspondant à un minimum global de la fonction $\mathcal{L}_{\text{ridge}}$. Commençons par remarquer que la fonction $\mathcal{L}_{\text{ridge}}$ est convexe en la variable $\mathbf{B}$; on peut donc trouver le point de minimum global en annulant la dérivée de $\mathcal{L}_{\text{ridge}}$ par rapport à $\mathbf{B}$. Cela conduit à l'équation suivante

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T). \tag{3.12}$$

Il s'agit de l'estimateur ridge du modèle de la régression linéaire de $\mathbf{Y} - \mathbf{U}\mathbf{V}^T$ par $\mathbf{X}$, en supposant que $\mathbf{U}$ et $\mathbf{V}$ sont connues.

Il faut maintenant minimiser la fonction

$$\mathcal{L}^{'}(\mathbf{U}, \mathbf{V}) = \mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \hat{\mathbf{B}}).$$

Considérons la décomposition en valeurs singulières de $\mathbf{X}$ telle que

$$\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T,$$

où $\mathbf{Q}$ est une matrice unitaire de taille $n \times n$, $\mathbf{R}$ une matrice unitaire de taille $d \times d$ et $\mathbf{\Sigma}$ une matrice de taille $n \times d$ contenant les valeurs singulières $\{\sigma_j\}_{j=1..d}$. L'écriture de $\mathcal{L}^{'}$ se simplifie comme ceci

$$\mathcal{L}^{'}(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\left\|\mathbf{D}_\lambda^2\mathbf{Q}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)\right\|_F^2 + \frac{1}{2}\lambda\left\|\mathbf{C}_\lambda\mathbf{Q}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)\right\|_F^2$$

où $\mathbf{C}_\lambda$ est une matrice de taille $d \times n$ remplie de zéro sauf sur la première diagonale qui contient les valeurs

$$\{\mathbf{C}_{\lambda,i,i}\}_{i=1..d} = \left\{\frac{\sigma_i}{\sigma_i^2 + \lambda}\right\}_{i=1..d}.$$

La matrice $\mathbf{D}_\lambda$ est une matrice diagonale de taille $n \times n$ contenant les termes

$$\{\mathbf{D}_{\lambda,i,i}\}_{i=1..n} = \left\{\sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, ..., \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, ..., 1\right\}.$$

Les matrices $\mathbf{D}_\lambda$ et $\mathbf{C}_\lambda$ étant diagonales, il est possible par le calcul de montrer que

$$\begin{aligned}\mathcal{L}^{'}(\mathbf{U}, \mathbf{V}) &= \frac{1}{2}\left\|\sqrt{(\mathbf{D}_\lambda^2 + \mathbf{C}_\lambda^2)}\mathbf{Q}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)\right\|_F^2 \\ &= \frac{1}{2}\left\|\mathbf{D}_\lambda\mathbf{Q}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)\right\|_F^2\end{aligned}$$

Enfin, optimiser la fonction objectif $\mathcal{L}^{'}$ équivaut au problème de trouver la meilleure approximation de rang K de la matrice

$$\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y},$$

qui est obtenue en tronquant la SVD pour ne garder que les K valeurs singulières les plus grandes (Eckart et Young, 1936). Nous avons bien montré que

$$\begin{aligned}\hat{\mathbf{U}}\hat{\mathbf{V}}^T &= \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \\ \hat{\mathbf{B}}^T &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T)\end{aligned}$$

est un point de minimum global de $\mathcal{L}_{\text{ridge}}$. □

3.3.3 Estimateur des moindres carrés régularisé en norme L_1

Dans cette partie, nous présentons un algorithme d'estimation des paramètres du modèle défini par (3.6) fondé sur un problème des moindres carrés régularisé en norme L_1 et en norme nucléaire. Nous montrons que cet algorithme calcule un minimum global du problème d'optimisation des moindres carrés régularisé.

Fonction objectif

Afin d'estimer les paramètres $\mathbf{U}$, $\mathbf{V}$ et $\mathbf{B}$ de LFMM, nous définissons la fonction objectif de type lasso suivante

$$\mathcal{L}_{\text{lasso}}(\mathbf{W}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \mu \left\| \mathbf{B} \right\|_1 + \gamma \left\| \mathbf{W} \right\|_*, \tag{3.13}$$

où $\mathbf{W}$ est la matrice latente définie en (3.8), $\left\| \mathbf{B} \right\|_1$ la norme L_1 de $\mathbf{B}$, μ le paramètre de régularisation L_1, $\left\| \mathbf{W} \right\|_*$ la norme nucléaire de la matrice $\mathbf{W}$, définie comme la somme de ses valeurs singulières, et γ le paramètre de régularisation de la norme nucléaire. Le choix de la norme L_1 est motivé par le fait que l'on s'attend à ce que seulement une certaine proportion des colonnes de $\mathbf{Y}$ soit associée à $\mathbf{X}$. Autrement dit, seules certaines lignes de la matrice des effets $\mathbf{B}$ doivent être non nulles. La régularisation L_1 est connue pour produire des estimateurs parcimonieux de $\mathbf{B}$ (Tibshirani, 1996). La fonction $\mathcal{L}_{\text{lasso}}$ fait aussi intervenir une régularisation sur la matrice latente $\mathbf{W}$. Nous ajoutons la régularisation en norme nucléaire afin de lever la contrainte de rang sur $\mathbf{W}$ empêchant de définir un problème d'optimisation convexe. Avec le terme de régularisation sur $\mathbf{W}$, la fonction $\mathcal{L}_{\text{lasso}}$ devient convexe. En outre, les pénalisations sur la norme nucléaire sont utilisées pour pénaliser le rang (Mishra et al., 2013). Ainsi, la régularisation en norme nucléaire contraint le rang de $\mathbf{W}$ et donc le nombre de variables latentes.

Algorithme de minimisation de la fonction objectif $\mathcal{L}_{\text{lasso}}$

Nous présentons maintenant un algorithme de descente par blocs de coordonnées qui permet d'estimer les paramètres de LFMM en minimisant la fonction objectif $\mathcal{L}_{\text{lasso}}$ définie par (3.13). Nous initialisons l'algorithme avec des matrices nulles :

$$\hat{\mathbf{W}}_{t=0} = 0$$
$$\hat{\mathbf{B}}_{t=0} = 0.$$

Nous alternons ensuite les deux étapes suivantes :

1. Calculer $\hat{\mathbf{B}}_t$ le point minimum de

$$\mathcal{L}'_{\text{lasso}}(\mathbf{B}) = \frac{1}{2}||(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T||_F^2 + \mu||\mathbf{B}||_1 \tag{3.14}$$

2. Calculer $\hat{\mathbf{W}}_t$ le point minimum de

$$\mathcal{L}''_{\text{lasso}}(\mathbf{W}) = \frac{1}{2}||(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T) - \mathbf{W}||_F^2 + \gamma||\mathbf{W}||_*. \tag{3.15}$$

Ces deux étapes sont répétées jusqu'à ce que l'algorithme converge ou bien que t atteigne le nombre maximum d'itérations fixé à l'avance. Nous allons maintenant expliquer plus en détail les deux étapes de l'algorithme décrites ci-dessus.

La première étape de l'algorithme consiste à faire une régression linéaire régularisée en norme L_1 de la matrice résiduelle

$$\mathbf{E}_t^1 = \mathbf{Y} - \hat{\mathbf{W}}_{t-1} \tag{3.16}$$

par les variables explicatives $\mathbf{X}$. Il existe plusieurs algorithmes pour estimer les paramètres de cette régression. Nous utilisons l'algorithme de descente par coordonnées de Friedman, Hastie, Höfling et al. (2007). Dans le cas présent, on s'intéresse plus particulièrement à l'estimation des variables latentes, qui permettront ensuite de faire le test d'association (voir la partie 3.3.6). Nous supposons donc que les variables explicatives $\mathbf{X}$ ont été transformées de sorte que

$$\mathbf{X}^T\mathbf{X} = \mathbf{Id}_d. \tag{3.17}$$

On a alors d'après Tibshirani (1996),

$$\hat{\mathbf{B}}_t = \text{sign}(\bar{\mathbf{B}}_t)(\bar{\mathbf{B}}_t - \mu)_+ \tag{3.18}$$

où

$$s_+ = \max(0, s), \tag{3.19}$$

$\text{sign}(s)$ est le signe de s et $\bar{\mathbf{B}}_t$ est l'estimateur du paramètre de la régression linéaire classique donné dans ce cas par

$$\bar{\mathbf{B}}_t = \mathbf{X}^T\mathbf{E}_t^1.$$

La deuxième étape de l'algorithme résout un problème d'approximation de rang faible de la matrice résiduelle

$$\mathbf{E}_t^2 = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T, \tag{3.20}$$

Cette approximation est donnée grâce à un seuillage des valeurs singulières de la matrice $\mathbf{E}_t^2$ (J.-F. Cai et al., 2010). Pour cela, on commence par calculer la décomposition en valeurs singulières de la matrice résiduelle :

$$\mathbf{E}_t^2 = \mathbf{MSN}^T, \tag{3.21}$$

où $\mathbf{M}$ est une matrice unitaire de taille $n \times n$, $\mathbf{N}$ une matrice unitaire de taille $p \times p$ et $\mathbf{S}$ une matrice de taille $n \times p$ contenant les valeurs singulières $\{s_j\}_{j=1..n}$. On a alors

$$\hat{\mathbf{W}}_t = \mathbf{M}\bar{\mathbf{S}}\mathbf{N}^T \tag{3.22}$$

où $\bar{\mathbf{S}}$ est la matrice diagonale formée par les valeurs singulières de $\mathbf{S}$ seuillées de sorte que

$$\bar{s}_j = (s_j - \gamma)_+, \ j = 1, ..., n.$$

Le seuillage produit des valeurs nulles et ramène vers zéro les valeurs singulières restantes.

L'algorithme de descente par blocs de coordonnées ne converge pas en général vers un point minimum quand la fonction objectif n'est pas continûment différentiable, comme c'est le cas pour la fonction $\mathcal{L}_{\text{lasso}}$. On peut trouver dans la littérature des résultats généraux sur les algorithmes par blocs de coordonnées dans des cas où la fonction objectif n'est pas différentiable (Tseng, 2001). Cependant, les théorèmes démontrés par Tseng (2001) dépassent largement le cadre de la convergence de l'algorithme d'estimation L_1 présenté ici et compliquent l'extraction des résultats intéressants. Pour faciliter la compréhension, nous proposons de démontrer un théorème plus faible qui s'applique directement à notre cas. Pour cela nous introduisons quelques notations. Soit la fonction f définie sur le domaine

$$A = A_1 \times A_2 \times ... \times A_m, \tag{3.23}$$

un produit cartésien d'ensembles fermés et convexes. L'algorithme de descente par blocs de coordonnées est défini par la formule de récurrence suivante :

$$x_i^{k+1} \in \arg\min_{\zeta \in X_i} f(x_1^k, ..., x_{i-1}^k, \zeta, x_{i+1}^k, ..., x_m^k), \ i = 1, ..., m. \tag{3.24}$$

En nous inspirant des résultats présentés par Tseng (2001) et de la proposition 2.7.1 de Bertsekas (1997) (ce dernier démontrant la convergence de l'algorithme de descente par blocs de coordonnées dans le cas où la fonction objectif est différentiable), nous pouvons énoncer le théorème suivant :

Théoreme 3. *Si f est une fonction continue de A dans $\mathbb{R}$, convexe et telle que*

$$f(x_1, ..., x_m) = g(x_1, ..., x_m) + \sum_{i=1}^{m} f_i(x_i), \tag{3.25}$$

où g est convexe et différentiable et les fonctions f_i sont continues et convexes. Soit $\{x^k\}$ la suite générée par (3.24). Alors tout point limite de $\{x^k\}$ est un point de minimum global de f.

Démonstration. On note

$$\bar{x} = (\bar{x}_1, ..., \bar{x}_m),$$

un point limite de $\{x^k\}$, la suite générée par (3.24) ; $\bar{x}$ est bien dans A le domaine de définition de f car cet ensemble est fermé. Comme g est convexe et différentiable on a pour tout $x \in A$

$$f(x) - f(\bar{x}) \geq \nabla g(\bar{x})(x - \bar{x}) + \sum_{i=1}^{m}(f_i(x_i) - f_i(\bar{x}_i)) \tag{3.26}$$

$$= \sum_{i=1}^{m}(\nabla_i g(\bar{x})(x_i - \bar{x}_i) + f_i(x_i) - f_i(\bar{x}_i)) \tag{3.27}$$

où $\nabla g(\bar{x})$ et $\nabla_i g(\bar{x})$ sont respectivement la dérivée et la dérivée par rapport à la i-ième variable de g en $\bar{x}$. D'autre part pour chaque variable d'indice i

$$\nabla_i g(\bar{x})(x - \bar{x}) + f_i(x_i) - f_i(\bar{x}_i) \geq (\nabla_i g(\bar{x}) + r_i)(x - \bar{x}) \tag{3.28}$$

où r_i est n'importe quelle sous-dérivée de la fonction convexe f_i en $\bar{x}_i$. Or nous savons par construction de $\bar{x}$ que

$$f(\bar{x}) \leq f(\bar{x}_1, ..., x_i, ..., \bar{x}_m), \ \forall x_i \in A_i. \tag{3.29}$$

Pour chaque variable x_i, on peut donc dire que zéro appartient à l'ensemble des sous-dérivées par rapport à la variable x_i de f en $\bar{x}_i$. On peut alors dire qu'il existe une sous-dérivée r_i telle que

$$\nabla_i g(\bar{x}) + r_i = 0. \tag{3.30}$$

Pour chaque variable d'indice i on a finalement

$$\nabla_i g(\bar{x})(x - \bar{x}) + f_i(x_i) - f_i(\bar{x}_i) \geq 0 \tag{3.31}$$

Finalement, en utilisant (3.31) et (3.14) nous avons

$$f(x) - f(\bar{x}) \geq 0, \ \forall x \in A. \tag{3.32}$$

□

Ce résultat démontre que l'algorithme d'estimation L_1 des paramètres du modèle LFMM converge vers un point de minimum global de $\mathcal{L}_{\text{lasso}}$.

3.3.4 Complexité des algorithmes

Dans cette partie nous abordons la complexité des algorithmes présentés dans les sections précédentes. On peut distinguer deux grandes étapes dans ces algorithmes. La première étape est le calcul de la décomposition en valeurs singulières, c'est-à-dire :

- le calcul de la matrice latente défini par l'équation (3.10) pour l'estimation L_2,
- la résolution du problème d'optimisation de la fonction $\mathcal{L}'_{\text{lasso}}$ définie par (3.15) pour l'estimation L_1.

La seconde étape est le calcul de la projection orthogonale sur l'espace engendré par les variables explicatives $\mathbf{X}$, c'est-à-dire :

- le calcul de la matrice des effets définie par l'équation (3.11) pour l'estimation L_2,
- la résolution du problème d'optimisation de la fonction $\mathcal{L}''_{\text{lasso}}$ définie par (3.14) pour l'estimation L_1.

D'après Halko et al. (2011), le calcul des K composantes dominantes de la décomposition en valeurs singulières demande $O(npK)$ opérations. Cette complexité peut être réduite à $O(np\log(K))$ opérations si on utilise une méthode avec projections aléatoires, comme celle présentée par Halko et al. (2011).

La deuxième étape importante consiste en une projection du résidu de l'approximation de rang faible sur l'espace engendré par $\mathbf{X}$. Le nombre précis d'opérations dépend des hypothèses qui sont faites sur la matrice $\mathbf{X}$. Dans l'algorithme d'estimation L_1, aucune inversion de matrice n'est nécessaire pour le calcul de $\hat{\mathbf{B}}_t$. Mais dans les deux algorithmes, si on s'intéresse seulement au comportement asymptotique par rapport à n, p et K, alors on peut majorer la complexité par $O(pn + K(p+n))$.

Finalement, pour les deux algorithmes, le nombre d'opérations est majoré par $O(npK)$. L'algorithme d'estimation L_1 est bien entendu plus long car il réalise plusieurs

fois les opérations de décomposition en valeurs singulières et de projection. L'algorithme d'estimation L_2 ne les réalise qu'une seule fois.

Outre la complexité temporelle il est important de garder à l'esprit la taille prise en mémoire, surtout pour ce genre d'algorithme qui prend en entrée des données potentiellement trop grandes pour la mémoire vive de l'ordinateur (RAM). Les algorithmes d'estimation L_1 et L_2 ne nécessitent pas de dupliquer la matrice des variables expliquées $\mathbf{Y}$. En effet, $\mathbf{Y}$ est de taille $n \times p$ et donc la dupliquer pourrait poser des problèmes sur des ordinateurs ne possédant pas assez de RAM. Il est possible d'envisager de ne pas charger $\mathbf{Y}$ en RAM et d'accéder aux données seulement quand cela est nécessaire.

3.3.5 Choix des hyperparamètres

La sélection des hyperparamètres est un problème commun à de nombreuses méthodes en analyse de données. Nous présentons plusieurs approches pratiques pour choisir les hyperparamètres qui interviennent dans les algorithmes que nous avons présentés ici. Nous commençons par présenter les différentes approches possibles pour choisir le nombre de variables latentes K. Nous présentons ensuite plusieurs heuristiques qui permettent d'aider au choix des paramètres de régularisation. Enfin nous présentons un algorithme de validation croisée adapté aux algorithmes que nous avons présentés.

Nombre de variables latentes (K)

Pour trouver le nombre de variables latentes K nous proposons d'isoler les variations de $\mathbf{Y}$ expliquées par les variables latentes à l'aide de la matrice $\mathbf{D}_\lambda$ utilisée dans l'estimation L_2 (voir la section 3.3.2). Pour cela on projette $\mathbf{Y}$ sur l'espace orthogonal à $\mathbf{X}$ en prenant $\lambda = 0$. On a alors

$$\mathbf{D}_0\mathbf{Q}^T\mathbf{Y} = \mathbf{D}_0\mathbf{Q}^T\mathbf{U}\mathbf{V}^T + \mathbf{D}_0\mathbf{Q}^T\mathbf{E}. \tag{3.33}$$

On peut ainsi utiliser les méthodes d'estimation du nombre K de variables latentes sur la matrice $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. Si on enlève les variables explicatives du modèle, les fonctions objectif des deux algorithmes d'estimations L_1 et L_2 correspondent à la fonction objectif de l'analyse en composantes principales (ACP). Nous utilisons donc les méthodes d'estimation du nombre de composantes dans l'ACP pour $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. Il existe

de nombreuses approches pour déterminer le nombre de composantes principales de l'ACP, très bien expliquées par Jolliffe (1986). On peut grouper les approches en trois catégories. La première catégorie regroupe les approches subjectives comme l'utilisation du scree plot (le graphe des valeurs singulières de la matrice des données). La seconde catégorie comprend les approches basées sur une modélisation de la distribution des données observées, comme par exemple la méthode présentée par Choi et al. (2014). La dernière catégorie est formée des approches basées sur la validation croisée, comme celle que nous détaillons plus loin. Aucune méthode ne s'est imposée comme la référence, et il est préférable d'en utiliser plusieurs. Pour les expériences que nous avons réalisées sur des données réelles, le choix du nombre de variables latentes K du modèle LFMM a été fait à partir du scree plot de l'ACP de la matrice $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. Nous avons aussi utilisé l'algorithme de validation croisée que nous présentons dans la section 3.3.5.

Paramètre de régularisation L_2

Le paramètre de régularisation L_2 (λ) intervient dans le calcul de l'estimation de la matrice latente décrit par l'équation 3.10. Ce paramètre de régularisation intervient dans la matrice diagonale $\mathbf{D}_\lambda$. Lorsque le paramètre de régularisation L_2 tend vers zéro, les variables $\mathbf{Y}$ et $\mathbf{X}$ sont linéairement décorrélées. Ainsi la matrice $\mathbf{D}_\lambda$ permet de réduire la corrélation entre les variables expliquées $\mathbf{Y}$ et les variables explicatives $\mathbf{X}$ afin de pouvoir estimer les variables latentes $\mathbf{U}$. Lorsque le paramètre λ tend vers l'infini, la matrice $\mathbf{D}_\lambda$ tend vers la matrice identité. L'estimation des variables latentes est alors calculée par l'ACP de la matrice $\mathbf{Y}$, sans prendre en compte la variable $\mathbf{X}$. Lorsque λ est trop grand, on risque d'expliquer par les variables latentes une partie de la variance de $\mathbf{Y}$, qui devrait être expliquée par $\mathbf{X}$, et donc manquer certaines associations. Le choix du paramètre de régularisation λ est une affaire de dosage, il ne doit être ni trop grand ni trop petit. Nous avons remarqué dans les expériences que de petites valeurs donnent de meilleurs résultats dans de nombreux cas.

Paramètre de régularisation L_1

Le paramètre de régularisation L_1 (μ) a un impact sur le nombre de lignes nulles dans la matrice des effets $\mathbf{B}$. La proportion de lignes non nulles dans la matrice $\mathbf{B}$ correspond à la proportion de colonnes de $\mathbf{Y}$ corrélées aux variables explicatives $\mathbf{X}$.

Quand on prend en compte les variables latentes plutôt que de choisir le paramètre de régularisation μ, il est équivalent de choisir la proportion de colonnes de $\mathbf{Y}$ corrélées aux variables $\mathbf{X}$. Pour trouver un paramètre de régularisation qui correspond à la proportion de lignes non nulles, nous proposons une heuristique basée sur un chemin de régularisation inspirée par les travaux de Friedman, Hastie et Tibshirani (2010). Nous commençons par la plus petite valeur du paramètre de régularisation μ tel que le vecteur

$$\hat{\mathbf{B}}_{t=1} = \text{sign}(\bar{\mathbf{B}}_{t=1})(\bar{\mathbf{B}}_{t=1} - \mu)_+ \tag{3.34}$$

vaut zéro. La matrice $\hat{\mathbf{B}}_{t=1}$ est le résultat de la première étape de l'algorithme d'estimation des moindres carrés régularisée en norme L_1 (cf section 3.3.3). La valeur de μ correspondante est notée $\mu^{\max}$. Ensuite, nous construisons une suite de m valeurs de μ décroissant selon une échelle logarithmique depuis $\mu^{\max}$ jusqu'à

$$\mu^{\min} = \epsilon\mu^{\max}. \tag{3.35}$$

Enfin, pour chaque valeur de la suite ainsi croissante, nous calculons le nombre de lignes non nulles dans $\hat{\mathbf{B}}$ et l'estimation de la matrice des effets. Nous stoppons l'algorithme si la proportion de lignes non nulles souhaitée est dépassée.

Paramètre de régularisation de la norme nucléaire

Le paramètre de régularisation de la norme nucléaire (γ) dans l'algorithme d'estimation L_1 a une influence sur le rang de la matrice latente $\mathbf{W}$. Il est préférable de choisir le rang de cette matrice, correspondant au nombre de variables latentes K, que de choisir le paramètre de régularisation γ. Nous proposons l'heuristique suivante pour calculer le paramètre γ à partir du nombre de facteurs K. Nous commençons par calculer les valeurs singulières de la matrice des variables explicatives $\mathbf{Y}$, notées $(\sigma_1, ..., \sigma_n)$. Ensuite, nous calculons

$$\gamma = \frac{(\sigma_K + \sigma_{K+1})}{2}. \tag{3.36}$$

Nous avons remarqué dans les expériences que ce choix du paramètre de régularisation γ a toujours fait converger l'algorithme d'estimation L_1 vers une estimation de la matrice latente $\hat{\mathbf{W}}$ de rang K.

Validation croisée

La validation croisée est une méthode d'évaluation des hyperparamètres d'un modèle, très utilisée en apprentissage statistique. Le principe est de séparer les données en une partie d'apprentissage et une partie de test. Les données d'apprentissage sont utilisées pour estimer les paramètres du modèle. On mesure ensuite l'erreur de prédiction à l'aide des données de test. Pour que la validation croisée fonctionne, il est très important que les données de test ne soient pas utilisées pour estimer les paramètres du modèle. Dans le cas des modèles à facteurs latents en général, les données d'apprentissage ne permettent toutefois pas de calculer les variables latentes pour les données de test. Une méthode consite à séparer les variables des données de test. Nous utilisons ensuite une partie des variables des données de test pour estimer les variables latentes et l'autre partie pour calculer l'erreur de prédiction (Bro et al., 2008). Nous présentons maintenant plus formellement notre procédure de validation croisée.

Nous commençons par séparer les données en une partie d'entraînement et une partie de test ; c'est-à-dire que nous séparons les matrices des variables expliquées $\mathbf{Y}$ et explicatives $\mathbf{X}$ en deux parties selon leurs lignes. Nous notons I l'ensemble des indices des lignes choisies pour estimer l'erreur de prédiction. On estime à partir des données d'entraînement la matrice des axes factoriels que l'on note $\hat{\mathbf{V}}_{-I}$ et la matrice des effets que l'on note $\hat{\mathbf{B}}_{-I}$. Ensuite, nous séparons les observations des variables expliquées de test en deux parties afin d'estimer les variables latentes sur les variables restantes. On notera J l'ensemble des colonnes de la matrice des variables expliquées $\mathbf{Y}$ sélectionnées pour estimer la matrice des variables latentes. La matrice des variables latentes est estimée de la manière suivante

$$\hat{\mathbf{U}}_{-J} = (\mathbf{Y}[I, -J] - \mathbf{X}[I,](\hat{\mathbf{B}}_{-I}[-J,])^T)\hat{\mathbf{V}}_{-I}[-J,]^T, \tag{3.37}$$

où l'opérateur crochet représente l'opérateur de sélection de lignes et colonnes d'une matrice. Enfin, on peut calculer l'erreur de prédiction comme ceci

$$\text{err} = \frac{1}{|I||J|} \left\| \mathbf{Y}[I, J] - \hat{\mathbf{U}}^{-J}\hat{\mathbf{V}}_{-I}[J,]^T - \mathbf{X}[I,]\hat{\mathbf{B}}_{-I}[J,]^T \right\|_F, \tag{3.38}$$

où $|I|$ est le nombre d'indices dans l'ensemble I. Cette procédure permet de mesurer une erreur sur des observations des variables expliquées qui n'ont pas été utilisées pour estimer les paramètres du modèle.

3.3.6 Tests d'hypothèse corrigés pour les facteurs de confusion

Jusqu'ici, nous avons abordé l'estimation des variables latentes et des effets. Cependant, l'objectif initial est de trouver les variables expliquées associées aux variables explicatives, tout en prenant en compte les variables latentes. Nous présentons dans cette partie un test d'hypothèse de nullité de l'effet, corrigé pour les variables latentes. Une approche simple consiste à considérer l'estimation des variables latentes $\hat{\mathbf{U}}$ comme les vraies valeurs de $\mathbf{U}$. Nous utilisons ensuite les variables latentes $\hat{\mathbf{U}}$ comme des variables explicatives dans le modèle. C'est une méthode très courante dans les études d'association qui a montré de très bons résultats quand il y suffisamment d'individus (Gerard et Stephens, 2017b ; Price et al., 2006 ; Song et al., 2015 ; Leek et Storey, 2008 ; Rahmani et al., 2016). Nous avons choisi de réaliser un test d'hypothèse qui repose sur la régression linéaire car cela correspond au modèle LFMM si on suppose que $\mathbf{U}$ est connue. Les estimations des variables latentes peuvent être traitées comme variables explicatives dans n'importe quel modèle statistique. On pourrait par exemple envisager d'utiliser un modèle de régression linéaire généralisée. Afin de simplifier les notations et sans perte de généralité, nous supposons qu'il n'y a qu'une seule variable explicative, c'est à dire que la dimension de la matrice $\mathbf{B}$, égale à d, vaut 1. De plus, nous signalons qu'il est possible d'ajouter d'autres variables explicatives à la régression. Cela a un intérêt si l'on observe des variables qui sont des facteurs de confusion connus pour notre étude d'association, comme par exemple l'âge et le genre des individus. Nous rappelons que l'estimation de la matrice des K variables latentes $\hat{\mathbf{U}}$ est définie de façon unique grâce à l'ACP de la matrice $\hat{\mathbf{W}}$. La matrice $\hat{\mathbf{W}}$ est estimée grâce aux algorithmes d'estimation L_1 ou L_2 de la matrice latente de LFMM que nous avons présentés précédemment.

Calcul de la statistique de test et des p-valeurs

Pour chaque variable expliquée $\mathbf{Y}_j$, nous avons défini le modèle de régression linéaire suivant

$$\mathbf{Y}_j = \hat{\mathbf{U}}\gamma_j^T + \mathbf{X}\beta_j + \mathbf{E_j}, \tag{3.39}$$

où la matrice $\hat{\mathbf{U}}$ est l'estimation de la matrice des variables latentes du modèle LFMM. Nous supposons que l'erreur $\mathbf{E}_j$ est gaussienne de moyenne nulle. Nous voulons tester l'hypothèse de nullité du coefficient de régression β_j. Sous ces hypothèses, nous pouvons

calculer pour chaque variable expliquée $\mathbf{Y}_j$ une statistique de test z_j, assimilable à un z-score. Sous l'hypothèse nulle, la statistique de test suit la loi de Student à $n - K - 1$ degrés de liberté. On peut donc calculer une p-valeur pour chaque variable expliquée $\mathbf{Y}_j$. Le détail du calcul de la statistique de test est très classique. Par exemple, il est donné dans la section 3.2 du livre de HASTIE et al. (2009).

Calibration des p-valeurs

Il arrive parfois que la statistique de test ne suive pas la distribution théorique sous l'hypothèse nulle. On dit dans ce cas que le test est mal calibré. EFRON (2004) propose des exemples de situations qui peuvent aboutir à des tests mal calibrés. Dans les études que nous présentons ici, on s'attend à ce que la majorité des variables expliquées ne soit pas associées à la variable d'intérêt. Ainsi une majorité des statistiques de test sont distribuées selon l'hypothèse nulle. Cela nous permet d'utiliser l'approche choisie par SUN et al. (2012), qui consiste à calculer la médiane et la déviation absolue à la médiane (MAD pour median absolute déviation) directement sur les statistiques de tests z. En effet, la médiane donne une estimation robuste de la moyenne et le MAD de l'écart-type. On a alors une nouvelle statistique de test

$$\tilde{z}_j = \frac{z_j - \text{median}(z_1, ..., z_p)}{\text{MAD}(z_1, ..., z_p)}. \tag{3.40}$$

Pour calculer les nouvelles p-valeurs, on suppose que $\tilde{z}_j$ suit une loi normale de moyenne nulle et d'écart type 1 sous l'hypothèse nulle.

3.3.7 Implémentation en R

Les deux nouvelles méthodes de test d'association avec correction pour les facteurs de confusion, que nous avons développées dans cette thèse, ont été implémentées dans le langage de programmation R. Nous les avons appelées respectivement lassoLFMM pour l'implémentation des estimateurs régularisés en norme L_1 et ridgeLFMM pour les estimateurs régularisés en norme L_2. Les algorithmes lassoLFMM et ridgeLFMM prennent en entrée la matrice $\mathbf{X}$ et la matrice $\mathbf{Y}$. Ils prennent également en entrée le nombre de variables latentes K. L'algorithme ridgeLFMM prend une valeur pour λ (le paramètre de régularisation L_2). L'algorithme lassoLFMM prend la proportion de lignes non nulles dans la matrice des effets $\mathbf{B}$. Enfin les deux algorithmes retournent

les estimations pour les paramètres de LFMM ainsi qu'une p-valeur pour le test d'association de chaque colonne de $\mathbf{Y}$ avec $\mathbf{X}$.

3.4 Comparaisons avec d'autres méthodes de correction pour les facteurs de confusion

Dans cette section nous présentons d'autres méthodes pour les études d'association avec et sans correction pour les facteurs de confusion. Celles-ci sont comparées aux méthodes lassoLFMM et ridgeLFMM dans la section suivante.

3.4.1 Régressions linéaire simple et avec les scores de l'ACP

Dans lassoLFMM et ridgeLFMM, les tests d'hypothèse utilisés pour détecter les associations reposent sur un modèle de régression linéaire de $\mathbf{Y}$ par $\mathbf{X}$ et sur l'estimation des facteurs latents $\bar{\mathbf{U}}$. Il est donc naturel de comparer nos algorithmes à la méthode reposant sur la régression linéaire de $\mathbf{Y}$ par $\mathbf{X}$. Dans ce cas, aucun facteur latent n'est pris en compte dans l'étude d'association. D'autre part, nous comparons nos tests à une méthode qui repose sur une estimation des variables latentes par l'ACP. Dans ce cas, il s'agit de faire une régression de $\mathbf{Y}$ par $\mathbf{X}$ et $\bar{\mathbf{U}}$. La matrice $\bar{\mathbf{U}}$) est définie comme la matrice des scores sur les K premières composantes principales. Un principe similaire est utilisé dans la méthode EIGENSTRAT (Price et al., 2006). Les deux méthodes, reposant sur la régression linéaire avec et sans les scores de l'ACP, ont été implémentées en langage R et nous les appelons respectivement lm et PCAlm.

3.4.2 Méthode "Surrogate Variable Analysis" (SVA) (Leek et Storey, 2007)

Il existe deux versions de SVA : sva-two-step (Leek et Storey, 2007) et sva-irw (Leek et Storey, 2008). La méthode sva-two-step se découpe en deux étapes : une étape d'estimation de la matrice des axes factoriels $\mathbf{V}$ et une étape d'estimation de la matrice des variables latentes $\mathbf{U}$. Lors de la première étape la méthode sva-two-step estime les axes factoriels en faisant une ACP de la matrice résiduelle de la régression linaire de $\mathbf{Y}$ par $\mathbf{X}$. En utilisant les notations de la section 3.3.2, cela correspond à faire

l'ACP de $\mathbf{D}_{(\lambda=0)}\mathbf{Q}^T\mathbf{Y}$. Ensuite la méthode sva-two-step détermine un sous-ensemble de colonnes de la matrice $\mathbf{Y}$ qui sont les moins corrélées à la variable $\mathbf{X}$. Le sous-ensemble de colonnes est utilisé pour estimer la matrice des variables latentes $\mathbf{U}$.

La deuxième version de SVA est itérative. Plutôt que d'estimer les variables latentes sur un sous-ensemble de colonnes de la matrice $\mathbf{Y}$, la méthode sva-irw attribue un poids à chacune d'entre elles. Sachant les variables latentes calculées à l'itération précédente, on calcule la probabilité que l'effet de la variable $\mathbf{X}$ sur chaque colonne de $\mathbf{Y}$ soit nul. Ensuite les probabilités sont utilisées pour attribuer un poids à chaque colonne de $\mathbf{Y}$ et une nouvelle estimation des variables latentes est calculée à l'aide d'une ACP qui prend en compte ces poids. La méthode itère ces deux étapes un nombre de fois choisi par l'utilisateur. Nous avons utilisé le package R sva fourni par les auteurs.

3.4.3 "High dimensional factor analysis and confounder adjusted testing and estimation" (CATE) (Wang et al., 2017)

Nous présentons dans cette partie la méthode cate (Wang et al., 2017). Pour faciliter les explications, nous considérons le cas où il n'y a qu'une variable explicative $\mathbf{X}$. Dans la méthode cate, on commence par transformer la matrice des variables expliquées $\mathbf{Y}$ afin d'isoler les variations expliquées par les facteurs latents. Pour effectuer la transformation, on applique une matrice de changement de base aux lignes de $\mathbf{Y}$ de sorte que le premier axe de la nouvelle base soit colinéaire à $\mathbf{X}$. Cette transformation permet d'avoir sur la première ligne de la matrice transformée les coefficients de la régression linéaire de $\mathbf{Y}$ par $\mathbf{X}$. Sur toutes les autres lignes nous avons le projeté orthogonal de $\mathbf{Y}$ par rapport à $\mathbf{X}$ qui correspond au résidu de la régression. La méthode cate utilise le projeté orthogonal de $\mathbf{Y}$ par rapport à $\mathbf{X}$ pour calculer les axes factoriels. Cette première étape est comparable à l'étape de calcul de la matrice $\mathbf{V}$ dans notre méthode ridgeLFMM (voir partie 3.3.2). Dans ridgeLFMM, plutôt que d'enlever complètement les variations de $\mathbf{Y}$ expliquées par $\mathbf{X}$, nous les réduisons en fonction du paramètre de régularisation λ. Comme cela a été montré par Wang et al. (2017), les méthodes sva et cate calculent la même matrice des axes factoriels. Cette matrice correspond à celle estimée par ridgeLFMM dans le cas où λ vaut zéro. La méthode cate diffère de sva dans sa façon de calculer les variables latentes et les effets

de $\mathbf{X}$ sur $\mathbf{Y}$. Les auteurs de cate ont modélisé explicitement la corrélation entre les variables explicatives et les variables latentes tel que

$$\mathbf{U} = \mathbf{X}\boldsymbol{\alpha}^T + \mathbf{Z} \tag{3.41}$$

où $\boldsymbol{\alpha}$ est la matrice des effets de la variable $\mathbf{X}$ sur les variables latentes $\mathbf{U}$ et $\mathbf{Z}$ est une matrice de bruit indépendant de $\mathbf{X}$. La matrice $\mathbf{Z}$ est estimée en même temps que la matrice des axes factoriels, grâce à l'ACP de la projection de $\mathbf{Y}$ sur l'orthogonal de $\mathbf{X}$. Pour estimer les effets corrigés pour les facteurs de confusion la méthode cate considère la régression linéaire de $\mathbf{Y}$ par $\mathbf{X}$ tel que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\tau}^T + \mathbf{E}, \tag{3.42}$$

où $\mathbf{E}$ est une matrice de bruit résiduel et $bm\tau$ est la matrice des effets de $\mathbf{X}$ sur $\mathbf{Y}$ non corrigés pour les facteurs de confusion. Afin de calculer les effets corrigé, la méthode cate utilise une régression linéaire robuste des effets non corrigés par les axes factorielles. Pour le comprendre, en injectant (3.41) dans l'équation (3.6), on peut écrire les effets de $\mathbf{X}$ sur $\mathbf{Y}$ non corrigé comme ceci

$$\boldsymbol{\tau} = \mathbf{B} + \mathbf{V}\boldsymbol{\alpha}^T, \tag{3.43}$$

où $\mathbf{B}$ est la matrice des effets corrigés dans l'équation (3.6). Enfin, la matrice des effets corrigés $\mathbf{B}$ est calculée comme le résidu de la régression linéaire robuste. La régression robuste permet d'enlever de l'estimation de $\boldsymbol{\alpha}$ les effets atypiques qui correspondent aux colonnes de $\mathbf{Y}$ associées à $\mathbf{X}$. Nous avons utilisé le package R cate fourni par les auteurs.

3.5 Expérimentations : données simulées et réelles

3.5.1 Données simulées à partir de données réelles

Nous avons simulé des données pour lesquelles nous connaissons les variables expliquées associées à la variable explicative. Pour cela, nous utilisons une matrice de variables expliquées $\mathbf{Y}$ d'un jeu de données réelles. Pour obtenir les données simulées, nous réalisons une analyse en composantes principales de la matrice $\mathbf{Y}$, et ne gardons

que les K premières composantes en fonction du nombre de facteurs de confusion que nous souhaitons simuler. Nous avons alors

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^T + \mathbf{E} \tag{3.44}$$

où $\mathbf{V}$ est la matrice des K premiers axes principaux et $\mathbf{U}$ est la matrice des scores calculés par l'ACP. La matrice $\mathbf{E}$ est une matrice d'erreur résiduelle. Nous simulons ensuite la variable explicative, $\mathbf{X}'$, ainsi que K variables latentes, $\mathbf{U}'$. Pour ce faire, les matrices $\mathbf{U}'$ et $\mathbf{X}'$ sont simulées à l'aide de la loi normale multidimensionnelle tel que

$$[\mathbf{U}\mathbf{X}] \sim \mathcal{N}(0, \mathbf{S}).$$

La matrice de covariance de la loi normale est choisie de sorte que $\mathbf{X}'$ et $\mathbf{U}'$ soient corrélées. Les coefficients de corrélation linéaire, notée c_k, de la variable $\mathbf{X}'$ avec variable latente $\mathbf{U}_k$, sont tirés selon une loi uniforme entre -1 et 1. Ensuite, afin de régler l'intensité de la corrélation entre les variables latentes et la variable explicative, nous multiplions les coefficients de corrélation par une valeur que l'on note ρ. La matrice de corrélation s'écrit

$$\mathbf{S} = \begin{bmatrix} s_1 & 0 & \cdots & \rho c_1 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & s_K & \rho c_K \\ \rho c_1 & \cdots & \rho c_K & 1 \end{bmatrix},$$

où s_k désigne la variance empirique de $\mathbf{U}_k$. Cette simulation permet de produire des variables latentes de même variance que les scores de l'ACP et une variable explicative corrélée aux variables latentes.

Enfin nous simulons une matrice des effets $\mathbf{B}'$, de sorte qu'une proportion fixée des lignes de $\mathbf{B}'$ soit non nulle et tirée selon une loi normale. La nouvelle matrice des variables expliquées est simulée de la façon suivante

$$\mathbf{Y}' = \mathbf{U}'\mathbf{V}^T + \mathbf{X}'\mathbf{B}'^T + \mathbf{E}. \tag{3.45}$$

Nous avons ainsi généré des données pour lesquelles nous connaissons les colonnes de $\mathbf{Y}'$ associées à la variable $\mathbf{X}'$. Elles correspondent aux lignes non nulles de la matrice $\mathbf{B}'$.

Le jeu de données réelles que nous avons choisi pour réaliser les simulations est issu de la base de données 1000Genome que nous présentons dans la partie 3.6.4

Consortium, 2015. Nous avons gardé seulement les chromosomes 1 et 2. Cela permet de simuler une matrice de 52211 variables expliquées pour 1758 individus. Nous avons choisi de simuler 5 variables latentes. Le coefficient d'intensité de la corrélation entre les variables latentes et la variable explicative prend une fourchette de valeurs entre 0.1 et 1. Cela a permis de simuler des jeux de données plus ou moins difficiles pour l'inférence des facteurs de confusion. Nous avons de plus choisi une proportion des variables expliquées associées à la variable explicative, entre 1% et 20%. Pour chaque paramètre de simulation, nous avons simulé 5 jeux de données, ce qui donne un total de 125 jeux de données.

Nous avons considéré une méthode oracle qui fait le test d'association entre $\mathbf{Y}$ et $\mathbf{X}$ en connaissant les variables latentes de la simulation. La méthode oracle utilise le test d'association présenté dans la partie 3.3.6 avec la matrice des variables latentes de la simulation. Ainsi la méthode oracle devrait donner les meilleurs résultats.

3.5.2 Mesure de comparaison des performances

Pour évaluer la capacité des méthodes à séparer les vrais associations des fausses association nous avons utilisé l'aire sous la courbe précision-rappel noté AUC. La précision est le nombre de vrai positif détecté par une méthode divisé par le nombre total de candidats. La précision est le complément à 1 du taux de fausse découverte. Le rappel est le nombre de vraies associations détectées divisé par le nombre total des vraies associations. Le rappel est parfois appelé puissance en statistique. Une méthode permettant de séparer parfaitement les vraies associations des fausse associations donne un AUC de 1.

Afin d'évaluer la calibration des p-valeurs renvoyées par les méthodes nous calculons le facteur d'inflation des p-valeurs attribuées aux variables vraiment non associées. Le facteur d'inflation est calculé comme la médiane des z-scores au carré divisée par le quantile à 50% de la loi du χ^2 à un degré de liberté (Devlin et Roeder, 1999). Si les p-valeurs des variables vraiment non associées sont réparties uniformément alors le facteur d'inflation vaut 1. Une méthode qui renvoie des p-valeurs correctement calibrées permet de calculer une liste de candidats avec un taux de fausses découvertes moyen contrôlé. Pour cela on peut par exemple utiliser l'algorithme de Benjamini-Hoshberg de Benjamini et Hochberg (1995) ou bien la q-valeur de Storey (2011).

3.5.3 Étude d'association entre des niveaux de méthylation de l'ADN et la polyarthrite rhumatoïde (EWAS)

La polyarthrite rhumatoïde est une maladie auto-immune d'origine inconnue. Dans cette étude nous souhaitons étudier le rôle de la méthylation de l'ADN dans le développement de la polyarthrite rhumatoïde. La méthylation de l'ADN est un processus au cours duquel un groupe méthyle est ajouté aux molécules d'ADN. La méthylation peut changer l'activité de l'ADN et en particulier modifier sa transcription en protéine. Pour cette étude, nous nous intéressons au niveau de méthylation de 485577 sites de l'ADN chez 354 individus atteints de polyarthrite rhumatoïde et chez 335 individus sains (Liu et al., 2013). Il est connu que la méthylation de l'ADN dépend de l'âge, du genre et de la consommation de tabac de chaque individu. Nous savons aussi que le type de cellule sur laquelle on prélève la mesure influence le niveau de méthylation. Tous ces facteurs peuvent être des facteurs de confusion pour l'étude d'association à la maladie. Ils ont été pris en compte explicitement dans les études d'association qui ont été faites à partir des mêmes données que celles étudiées ici (Rahmani et al., 2016 ; Zou et al., 2014). Afin d'évaluer la capacité des méthodes à bien corriger les tests d'association, nous ne prenons pas en compte les facteurs de confusion connus et nous comparons les résultats à ceux obtenus par les études réalisées par Rahmani et al. (2016) et Zou et al. (2014) ; études qui prennent en compte explicitement les facteurs de confusion connus (type cellulaire, genre, consommation de tabac et âge).

De la même façon que Zou et al. (2014), nous avons filtré les sites avec un niveau de méthylation moyen constitutif, c'est-à-dire inférieur à 0.2 ou supérieur à 0.8. De plus, nous avons centré et divisé par l'écart-type les données de méthylation. Nous avons ensuite appliqué les méthodes cate, lm, PCAlm, sva-irw, sva-two-step, lassoLFMM et ridgeLFMM afin de trouver les sites de méthylation de l'ADN associés à la polyarthrite.

Enfin pour chaque méthode, nous avons calculé la liste obtenue lorsque l'on contrôle le taux de fausse découverte (FDR) à 1%. Les algorithmes de contrôle du FDR nécessitent que les p-valeurs soient correctement calibrées. Pour cela nous avons calibrées les p-valeurs grâce à la méthode présentée dans la partie 3.3.6. Le contrôle du FDR a été réalisé à l'aide du package R `qvalue` (Storey, 2011).

3.5.4 Étude d'association entre des données génétiques et la maladie cœliaque (GWAS)

La maladie cœliaque est une maladie auto-immune ayant une prévalence de près de 1% dans la population générale (Gujral, 2012). Bien que les mécanismes d'apparition de cette maladie ne soient pas compris, des études montrent de fortes associations avec certains gènes (Dubois et al., 2010), laissant envisager des causes génétiques à la maladie. Comme la maladie cœliaque est très étudiée, faire une étude d'association de celle-ci avec des données génomiques constitue un bon test pour les méthodes de correction des facteurs de confusion. Nous pourrons en effet comparer nos résultats à ceux des nombreuses autres GWAS de la maladie cœliaque. Pour cela, nous avons utilisé la base de données GWAS catalog pour récupérer les SNPs ayant été identifiés dans d'autres études comme étant associés avec la maladie cœliaque (MacArthur et al., 2016). Par ailleurs, nous savons que la stratification des individus en populations peut être un facteur de confusion dans les GWAS. Une pratique commune consiste à corriger les GWAS en utilisant les scores de l'ACP des données génétiques (Price et al., 2006). Nous proposons ici de faire une étude d'association entre la maladie cœliaqueet des données génétiques présentées dans (Dubois et al., 2010). Ces données comportent 281122 SNPs pour 15155 individus, 10659 témoins et 4496 cas.

Avant l'étude d'association, nous avons filtré les données génétiques afin de garder seulement les SNPs ayant le variant le moins fréquent présent dans au moins 5% des observations. Nous avons de plus filtré les individus trop apparentés. Pour cela, nous avons mesuré la probabilité qu'une séquence de SNPs, prise chez deux individus différents, soit identique. Si la probabilité est supérieure à 0.08, nous n'avons gardé qu'un des deux individus. Par ailleurs, il est connu que les SNPs sont corrélés le long du génome par liaison génétique (LD, linkage disequilibrium). Nous avons donc filtré les SNPs très corrélés entre eux. Cette étape est identifiée dans la littérature comme l'étape de *LD pruning*. Pour chaque SNP nous n'avons gardé que le SNP de variance maximum sur une fenêtre de 100 SNPs, et cela sur les SNPs ayant un coefficient de corrélation linéaire au carré qui est supérieur à 0.2. Cette procédure de prunning nous a permis d'identifier un sous-ensemble de 80275 SNPs. Les opérations de filtrage ont été effectuées à l'aide du logiciel `plink` (Purcell et al., 2007). Enfin nous avons utilisé le logiciel beagle pour imputer les données manquantes de la matrice de SNPs

(B. L. Browning et S. R. Browning, 2016).

Nous avons appliqué les méthodes cate, lm, PCAlm, lassoLFMM et ridgeLFMM dans le but de trouver les SNPs associés à la maladie cœliaque. Afin d'estimer les facteurs de confusion, nous avons appliqué les méthodes sur le sous-ensemble de 80275 SNPs identifiés par l'étape de pruning. Par la suite, nous avons effectué le test d'hypothèse présenté dans la section 3.3.6 sur les 281122 SNPs, en utilisant l'estimation des variables latentes calculées sur le sous-ensemble de 80275 SNPs. Les 281122 SNPs ont finalement été groupés par voisinage de SNPs corrélés. Cette étape a été réalisée à l'aide de l'algorithme de *clumping* du logiciel `plink`. La p-valeur attribuée au groupe est la plus faible p-valeur du groupe. Comme dans l'étude EWAS, le contrôle du FDR a été réalisé à l'aide du package R `qvalue`.

3.5.5 Étude d'association entre des données génétiques et un gradient environnemental (GEAS)

Depuis Darwin nous savons que les organismes vivants s'adaptent à leur environnement (Darwin, 1859). Ainsi les organismes les mieux adaptés à leur habitat ont plus de chance de survivre et de se reproduire. Si l'avantage adaptatif a des causes génétiques, c'est-à-dire qu'il existe une combinaison de gènes qui confère à l'organisme une fonction lui permettant d'être mieux adapté à son habitat, alors le patrimoine génétique qui confère l'avantage adaptatif est transmis aux futures générations. Nous pouvons chercher à détecter les signatures génétiques laissées par l'adaptation à l'environnement, comme nous l'avons fait dans le chapitre précédent.

Nous proposons maintenant de détecter les signatures de l'adaptation à l'environnement en faisant une étude d'association d'un gradient environnemental avec des données génétiques. Pour cela nous avons récupéré les données génétiques du projet 1000Genome phase 3 (Consortium, 2015). Ces données regroupent 84.4 millions de variants génétiques pour 2506 individus venant de 26 populations différentes. Par ailleurs nous avons récupéré des données climatiques à partir de la base de données WorldClim 2.0 (Fick et Hijmans, 2017). Nous avons filtré les individus trop apparentés et les variants génétiques de faible fréquence, de la même façon que l'étude GWAS (voir section 3.5.4). Nous avons enlevé les populations afro-américaine et africaine des Caraïbes pour ne garder que les individus vivant dans leur environnement depuis plusieurs générations. Après cette étape de filtrage il reste 1409 individus et 5397214

locus. Enfin, de la même façon que pour l'étude GWAS, nous avons identifié un sous-ensemble de 296948 SNPs grâce au *LD pruning*. Nous avons utilisé ce sous-ensemble de locus pour calculer les variables latentes avec les méthodes PCAlm, cate, lassoLFMM et ridgeLFMM. Nous avons calculé ensuite des p-valeurs pour chacun des 5397214 locus et pour chacune des méthodes. Les candidats retournés par les méthodes a été étudiée grâce au logiciel VEP (Variant Effect Predictor). Le logiciel VEP permet d'annoter l'effet d'un polymorphisme sur l'expression des gènes (McLaren et al., 2016). Les annotations retournées par VEP sont classées en catégories d'importance : *LOW*, *MODERATE* et *HIGH*. Nous avons réalisé un test exact de Fisher afin d'étudier la sur-représentation de chacune des catégories d'importance. Nous avons enfin utilisé le package BiomaRt afin d'annoter les SNPs qui ont été associés à un phénotype dans d'autres études.

Afin de calculer le gradient climatique utilisé pour l'étude d'association, nous avons attribué une position géographique à chaque individu en prenant la capitale de son pays d'origine. Cela nous a permis de récupérer les données climatiques pour les positions géographiques à partir de la base de données WorlClim. Pour définir une variable explicative, nous n'avons gardé que la première composante principale des variables de la base Wordclim. La Figure 3.1 représente la valeur du gradient climatique pour chaque population.

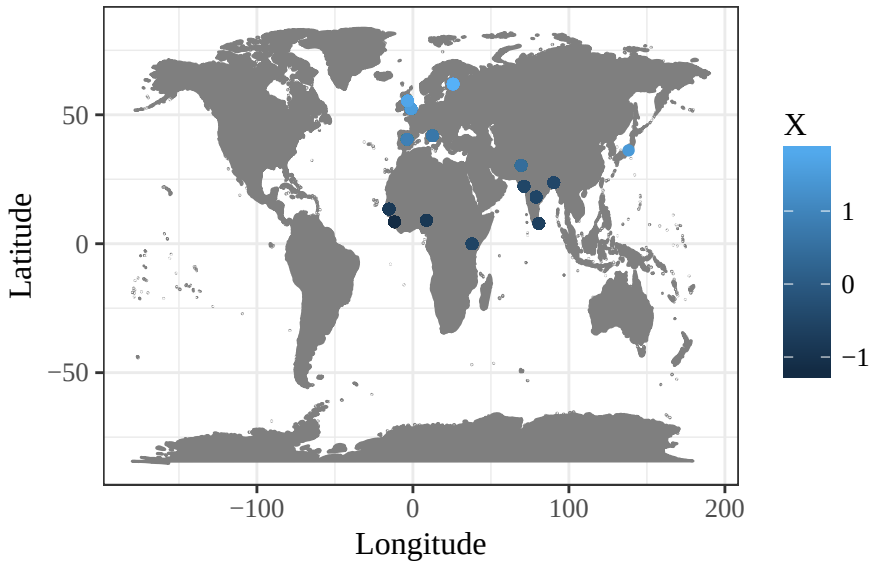

Figure 3.1 – Gradient climatique **X** utilisé pour l'association génotypes-environnement.

3.6 Résultats

Dans cette partie, nous présentons les expériences numériques que nous avons réalisées, pour évaluer la performance de nos algorithmes de correction des facteurs de confusion dans les études d'association. Les deux nouveaux algorithmes ont été implémentés dans le langage de programmation R et sont appelés respectivement ridgeLFMM pour l'estimation L_2 et lassoLFMM pour l'estimation L_1. Nos méthodes sont comparées aux méthodes présentées dans la partie 3.4. Les méthodes de régression linéaire avec et sans les scores de l'ACP ont été implémentées dans le langage R. Elles sont respectivement appelées lm et PCAlm. Pour les méthodes cate, sva-irw et sva-two-step, nous avons respectivement utilisé les implémentations R mises à disposition par leurs auteurs respectifs.

3.6.1 Comparaison des méthodes sur des données simulées

Sur les 125 jeux de données simulés, nous avons utilisé les méthodes lm, PCAlm, sva-irw, sva-two-step, cate, les deux méthodes présentées dans ce chapitre (lassoLFMM et ridgeLFMM) et la méthode oracle. Pour comparer les méthodes, nous avons calculé les facteurs d'inflation et les aires sous la courbe précision-rappel (AUC). Les méthodes cate, lassoLFMM et ridgeLFMM ont des performances similaires à la méthode oracle sur toutes les simulations (Figure 3.2), sauf lorsque paramètre de corrélation ρ vaut 1. En effet, sur les simulations avec une forte corrélation entre les variables latentes et la variable explicative, cate et ridgeLFMM renvoient des p-valeurs avec un taux d'inflation moyen de 3.3 alors que celui de lassoLFMM vaut 1.3 et celui de l'oracle 1.0 (Figure 3.2 D). Nous constatons par ailleurs que les méthodes cate et ridgeLFMM donnent des résultats très similaires sur toutes les simulations. Les performances de la méthode lm sont sensibles au paramètre de corrélation ρ. Lorsque le paramètre de corrélation est faible, l'AUC de lm et le facteur d'inflation de lm sont proches de ceux de la méthode oracle ($\rho = 0.1$, Figure 3.2 B et D). Mais le facteur d'inflation de lm croît jusqu'à plus de 30 et l'AUC décroît jusqu'à la moitié de celui de la méthode oracle lorsqu'il y a une forte corrélation entre les variables latentes et la variable explicative ($\rho = 1$, Figure 3.2 B et D). Les p-valeurs de la méthode PCAlm sont toujours correctement calibrées puisque le facteur d'inflation est toujours autour de 1 (Figure 3.2 C et D). Cependant l'écart de l'AUC de PCAlm avec l'AUC obtenu par

la méthode oracle croît avec la proportion de vraies associations et le paramètre de corrélation ρ (Figure 3.2 A et B). Les méthodes sva-two-step et sva-irw renvoient des p-valeurs correctement calibrées sauf quand le paramètre de corrélation ρ vaut 0.8 et 1.0 (Figure 3.2 C et D). L'AUC de sva-irw est toujours en dessous de l'AUC de la méthode oracle pour toutes les proportions de vraies associations dans les simulations (Figure 3.2 A). La différence entre l'AUC de sva-irw et l'AUC de la méthode oracle croît avec le paramètre de corrélation ρ (Figure 3.2 B). Nous observons également que l'AUC de sva-two-step est très légèrement en dessous de l'AUC de la méthode oracle pour toutes les proportions de vraies associations dans les simulations (Figure 3.2 A) et, comme pour sva-irw, la différence de l'AUC de sva-two-step avec l'AUC de la méthode oracle croît avec le paramètre de corrélation ρ mais plus faiblement que pour sva-irw (Figure 3.2 B).

3.6.2 Étude d'association entre des niveaux de méthylation de l'ADN et la polyarthrite rhumatoïde (EWAS)

Nous avons appliqué les méthodes cate, lm, PCAlm, sva-irw, sva-two-step, lassoLFMM et ridgeLFMM afin de trouver les sites de méthylation de l'ADN associés à la polyarthrite. Nous avons choisi $K = 10$ pour le nombre de variables latentes (voir Figure 3.3 A et B). Pour $K = 10$, les valeurs du paramètre de régularisation L_2 (λ) entre 10^{-10} et 1.0 donnent les mêmes valeurs moyennes d'erreur de prédiction (Figure 3.3 C). Nous avons choisi de prendre $\lambda = 10^{-5}$ pour ridgeLFMM.

Nous constatons une augmentation du nombre de petites p-valeurs pour toutes les méthodes ; cette augmentation est encore plus forte pour les méthodes lm et sva-irw (Figure 3.4 A). La figure 3.4 B montre la proportion des candidats identifiés par Zou et al. (2014) et Rahmani et al. (2016) qui sont retrouvés dans les listes rangées (les listes sont rangées selon les p-valeurs) retournées par chaque méthode. Nous rappelons que les candidats identifiés par Zou et al. (2014) et Rahmani et al. (2016) sont au nombre de 5. Ils ont été identifiés en prenant en compte les facteurs de confusion tels que l'âge, le sexe et une estimation de la composition cellulaire. Les méthodes considérées dans notre analyse retrouvent les candidats de la littérature dans leurs 40 premiers rangs, sauf lm et sva-irw (Figure 3.4 B). Pour la méthode lm, il faut prendre le rang 11881 pour trouver le premier candidat de Zou et al. (2014) et Rahmani et al. (2016), et le rang 138038 pour tous les avoir. Pour la méthode sva-irw, les candidats

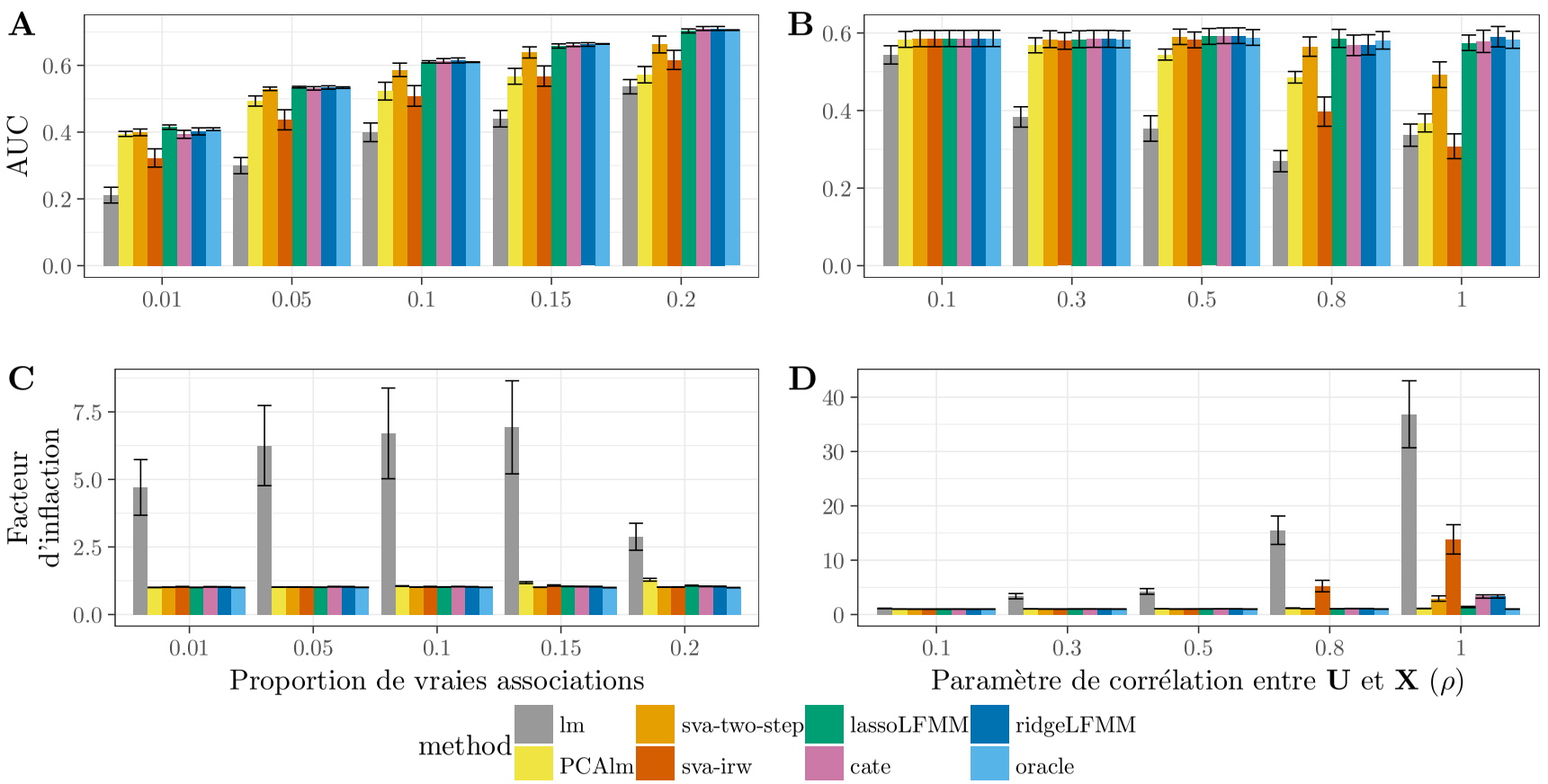

FIGURE 3.2 – **Comparaison des méthodes sur des simulations rélisées à partir du jeu de données 1000Genomes.** A-B) Aire sous la courbe précision-rappel en fonction de la proportion de colonnes de **Y** associées à **X** et de la corrélation de la variable explicative **X** avec les variables latentes. C-D) Facteur d'inflation calculé sur les variables nulles en fonction de la proportion de colonnes de **Y** associées à **X** et la corrélation de la variable explicative **X** avec les variables latentes.

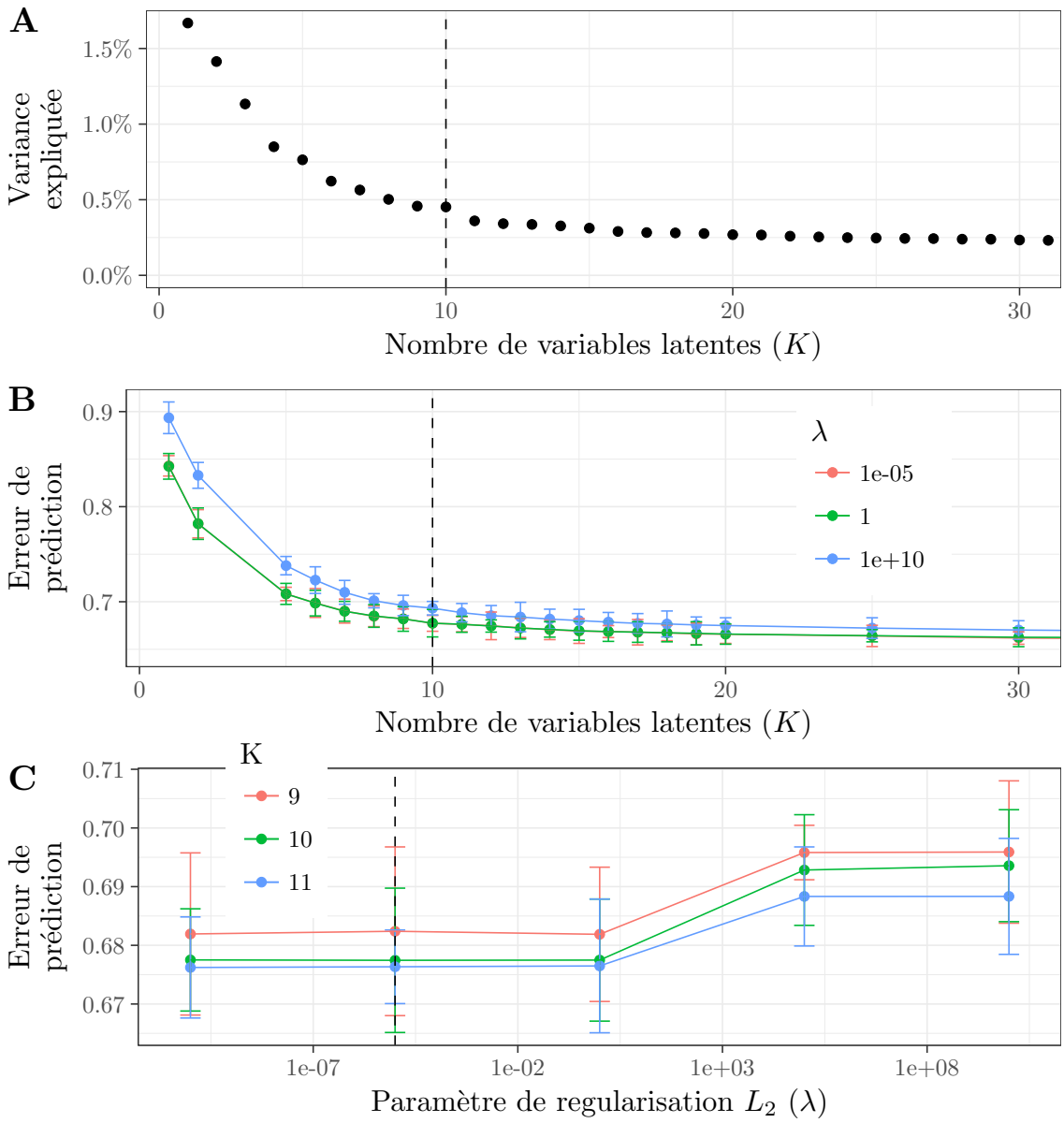

FIGURE 3.3 – **Choix des paramètres pour l'étude d'association entre des sites méthylation de l'ADN et la polyarthrite rhumatoïde (EWAS).** A) Proportion de variance expliquée de la projection de $\mathbf{Y}$ sur l'espace orthogonal à $\mathbf{X}$ (c'est à dire $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$) par chacune des composantes principales. B)C) Erreur de prédiction calculée grâce à la validation croisée des estimateurs L_2 des paramètres de LFMM pour différente valeurs du paramètre de régularisation λ et du nombre variable latentes K. Le point représente l'erreur de prédiction moyen et les barres l'erreur standard. La ligne pointillée verticale marque sur A et B le nombre de variables latentes choisies, c'est à dire 10, et sur C le paramètre de régularisation choisi, c'est à dire $\lambda = 10^{-5}$.

de Zou et al. (2014) et Rahmani et al. (2016) sont tous identifiés entre les rangs 5111 et 87659.

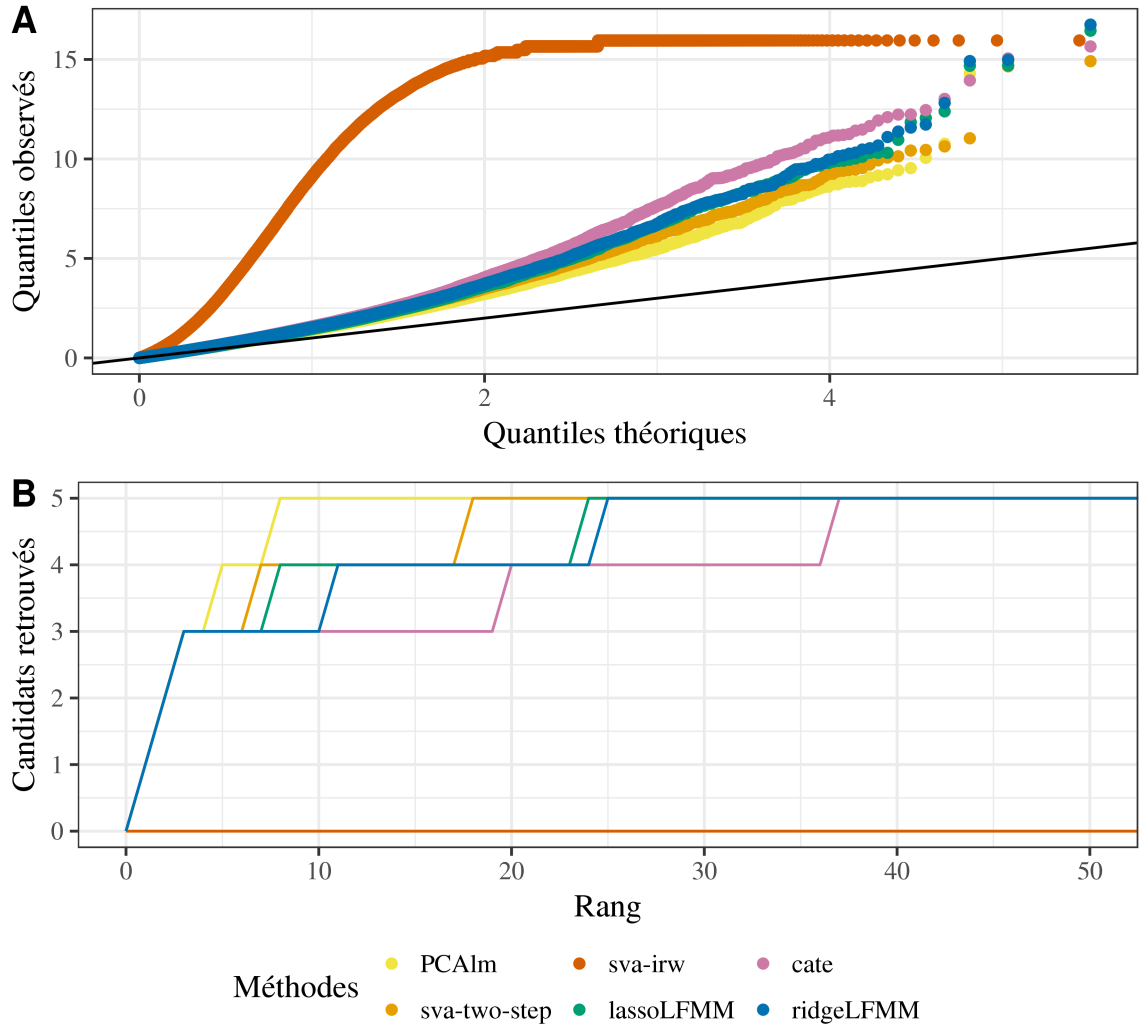

Figure 3.4 – **Q-Q plot et rang pour l'étude d'association entre des sites de méthylation et la polyarthrite rhumatoïde (EWAS).** A) Diagramme quantile-quantile de l'inverse du logarithme en base 10 des p-valeurs renvoyées par chaque méthode. Les quantiles théoriques sont ceux de la loi exponentielle. B) Nombre de sites proposés par Rahmani et al. (2016) et Zou et al. (2014) retrouvés dans la liste rangée (la liste est rangée selon les p-valeurs) avant un rang donné.

Enfin, nous calculons les sites de méthylation candidats pour l'association à la polyarthrite lorsque l'on contrôle le taux de fausse découverte à 1%. Nous avons écarté lm et sva-irw car ils renvoyaient des listes trop différentes des autres méthodes. Parmi les 19 candidats trouvés par toutes les méthodes lorsque nous contrôlons le FDR à

1%, nous retrouvons les 5 candidats identifiés par Zou et al. (2014) et Rahmani et al. (2016). Avec un FDR controlé à 1%, les méthodes PCAlm et sva-two-step retournent des listes de 20 candidats avec 19 candidats en commun. Les méthodes cate, lassoLFMM et ridgeLFMM proposent des listes de candidats plus de deux fois plus longues que les méthodes PCAlm et sva-two-step. De plus, les méthodes cate, lassoLFMM et ridgeLFMM proposent en commun 51 sites de méthylation (Figure 3.5). Enfin, nous remarquons que d'autres sites de méthylation situés dans des gènes sont détectés avant les 5 candidats de la littérature. En particulier, nous trouvons des gènes situés dans le complexe de gènes HLA qui joue un rôle central dans le système immunitaire (Table 3.1).

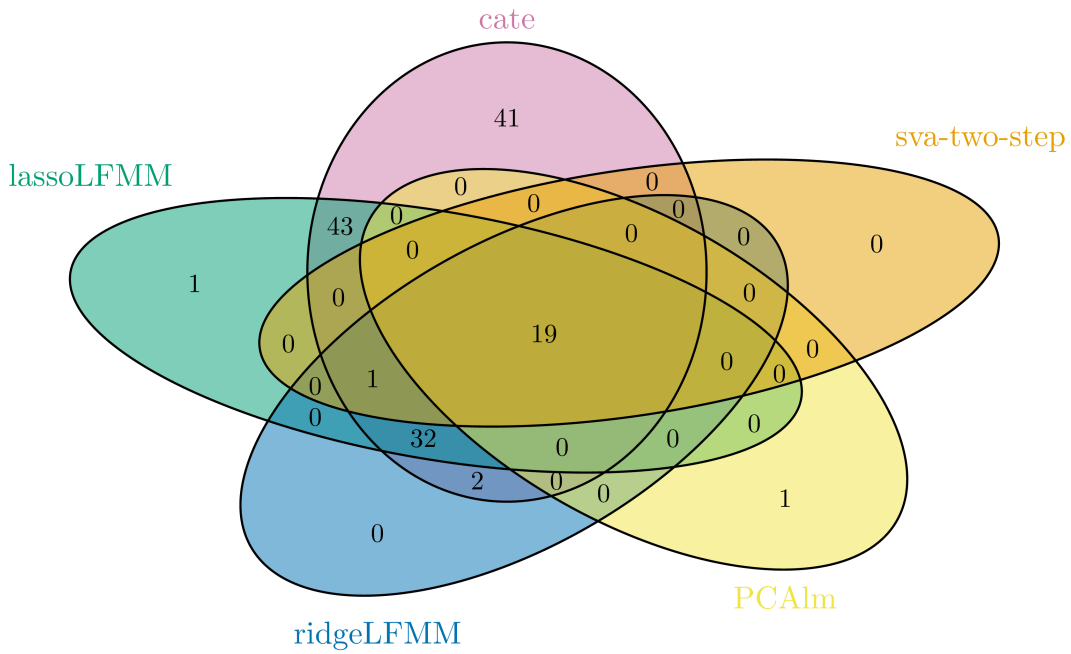

Figure 3.5 – **Diagramme de Venn de l'étude d'association entre des sites de méthylation et la polyarthrite rhumatoïde (EWAS).** Diagramme de Venn des listes controlées à un taux de fausses de découvertes de 1 %.

3.6.3 Étude d'association entre des données génétiques et la maladie cœliaque (GWAS)

Nous avons utilisé les méthodes cate, lm, PCAlm, lassoLFMM et ridgeLFMM afin de trouver les SNPs associés avec la maladie cœliaque. Nous avons utilisé les méthodes

avec 9 variables latentes (Figure 3.6 A et B). La validation croisée du paramètre λ tend à sélectionner une forte valeur pour celui-ci (Figure 3.6 C). La validation croisée présentée sur la Figure 3.6 C sélectionne une valeur plus grande que 10^3. Nous avons choisi 10^3 car, comme nous l'avons discuté dans la partie 3.3.5, λ trop grand conduit la méthode ridgeLFMM à renvoyer les mêmes résultats que PCAlm.

Les méthodes fournissent des p-valeurs bien calibrées avec une forte inflation du nombre de petites p-valeurs (Figure 3.7 A). La méthode la plus libérale est lm, alors que ridgeLFMM et PCAlm sont les méthodes les plus conservatives. Le chromosome 6 contient le complexe de gènes HLA. Ce complexe de gènes joue un rôle important dans le système immunitaire et dans la maladie cœliaque (Dubois et al., 2010). Nous avons séparé les SNPs candidats du GWAS catalog retrouvés sur le chromosome 6 des autre chromosomes. Les candidats du GWAS catalog sur le chromosome 6 arrive plus tôt dans les listes rangées (la liste est rangée selon les p-valeurs) des méthodes PCAlm et ridgeLFMM (Figure 3.7 B). Sur les autres chromosomes, ce sont les méthodes lassoLFMM et cate qui retrouvent le plus vite les candidats du GWAS catalog (Figure 3.7 B).

Enfin, nous calculons les SNPs candidats pour l'association à la maladie cœliaque-lorsque l'on contrôle le taux de fausse découverte à 1%. Ce sont ridgeLFMM et PCAlm qui donnent les plus petites listes contrôlées avec 754 candidats pour ridgeLFMM et 777 candidats pour PCAlm (Figure 3.8). Les méthodes cate et lm donnent les plus grandes listes avec le même nombre de 1319 candidats. Nous constatons que l'intersection des listes à FDR contrôlé de toutes les méthodes contient 28% des candidats du GWAS catalog. Afin d'identifier des groupes de SNPs associés à la maladie cœliaque, nous avons regroupé les SNPs voisins corrélés grâce à l'algorithme de *clumping* du logiciel `plink`. Dans les 20 premières zones détectées par lassoLFMM, en dehors du chromosome 6, 6 zones ne sont pas référencés dans le GWAS catalog (Table 3.2).

3.6.4 Étude d'association entre des données génétiques et un gradient environnemental (GEAS)

Nous avons utilisé les méthodes cate, ridgeLFMM, lassoLFMM et PCAlm avec 9 variables latentes, afin de trouver les SNPs associés à un gradient climatique. Nous avons utilisé les méthodes avec 9 variables latentes (Figure 3.9 A et B). Le screeplot et l'erreur de validation croisée conduisent à choisir 5 variables latentes,

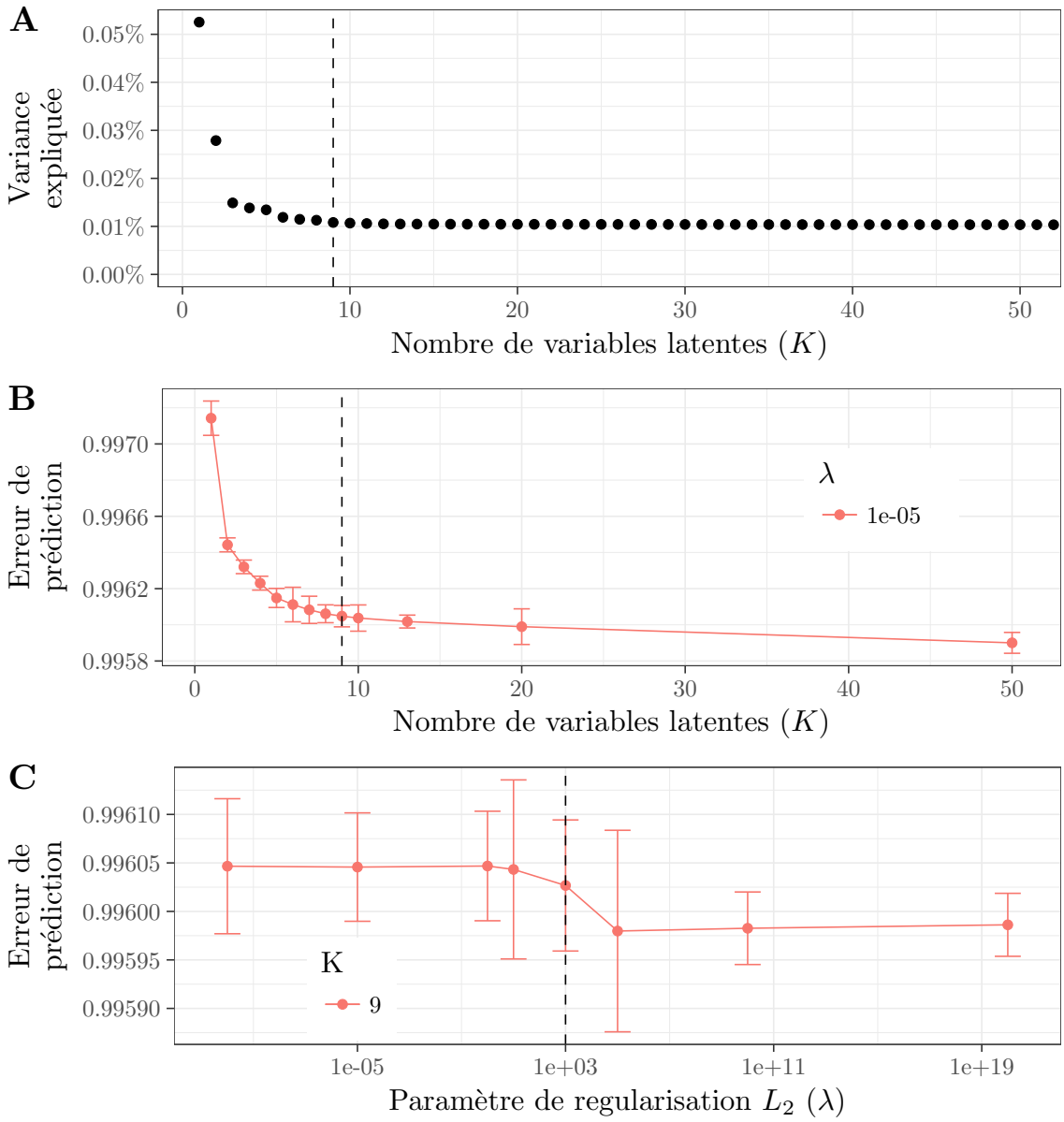

FIGURE 3.6 – **Choix des paramètres pour l'étude d'association entre génotype et la maladie cœliaque (GWAS).** A) Proportion de variance expliquée de la projection de $\mathbf{Y}$ sur l'espace orthogonal à $\mathbf{X}$ (c'est à dire $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$) par chacune des composantes principales. B)C) Erreur de prédiction calculée grâce à la validation croisée des estimateurs L_2 des paramètres de LFMM pour différente valeurs du paramètre de régularisation λ et du nombre variable latentes K. Le point représente l'erreur de prédiction moyen et les barres l'erreur standard. La ligne pointillée verticale marque sur A et B le nombre de variables latentes choisies, c'est à dire 9, et sur C le paramètre de régularisation choisi, c'est à dire $\lambda = 10^{-3}$.

mais nous avons choisi un nombre légèrement plus élevé car les individus proviennent de 16 populations différentes. En effet, les 6 premières variables latentes permettent d'expliquer la structure de population (Figure 3.10 A et B). Les variables latentes 8 et 9 ne permettent pas de visualiser la structure de population, mais il pourrait s'agir de facteurs de confusion à prendre en compte dans l'étude d'association (Figure 3.10 C). Enfin, les 10-ième et 11-ième variables latentes séparent un seul individu du reste du groupe (Figure 3.10 D), nous avons donc choisi $K = 9$. Pour $K = 9$, les valeurs du paramètre de régularisation L_2 comprises entre 10^{-10} et 1.0 donnent les mêmes valeurs moyennes d'erreur de prédiction (Figure 3.9 C). Nous avons choisi de prendre $\lambda = 10^{-5}$ pour ridgeLFMM.

La figure 3.11 montre la distribution observée des p-valeurs renvoyées par chaque méthode, contre la distribution théorique sous l'hypothèse nulle. Nous constatons une forte inflation pour les petites p-valeurs de la méthode lm. En effet, le MAD des z-scores retournées par la méthode lm est de 8.7. De plus nous remarquons que le diagramme quantile-quantile de lm forme une droite. Cela signifie que même si nous recalibrons les p-valeurs retournées par la méthode lm, celle-ci ne permet pas d'identifier des p-valeurs atypiques et donc des candidats pour l'association. Par contre, nous observons un excès de p-valeurs atypiques pour les autre méthodes. Pour les méthodes PCAlm, ridgeLFMM, lassoLFMM et cate, nous avons calculé la liste des SNPs candidats pour l'association au gradient climatique, lorsque nous contrôlons le FDR à 1%. La méthode lm a été écartée pour les raisons que nous venons d'évoquer. La Figure 3.8 montre l'intersection des listes de candidats contrôlés à un FDR de 1% entre les méthodes. L'union des candidats retournés par ces quatre méthodes donne 836 SNPs. Nous avons étudié la sur-représentation de chacun des degrés d'annotation de vep dans la sous-liste des 836 SNPs par rapport à la liste complète de SNPs. Nous constatons qu'il y a en proportion 22 fois plus d'annotation *HIGH*, 8 fois plus de *MODERATE* et 1.7 fois plus de *LOW* dans les 836 SNPs que parmi les 5397214 SNPs. Nous avons de plus testé si chaque rapport de proportion est significativement supérieur à 1 à l'aide d'un test exact de Fisher ; les trois tests renvoient une p-valeur inférieure à 0.005. Ainsi, parmi les SNPs que nous proposons pour l'association aux conditions climatiques, il y a une sur-représentation significative de polymorphismes ayant des impacts forts sur l'expression des gènes. Enfin, dans l'union des candidats, nous observons 21 SNPs référencés dans le GWAS catalog pour être associés à différents phénotypes liés à

l'environnement (Table 3.3). Nous remarquons qu'aucun SNP référencé dans le GWAS catalog n'est détecté par les méthodes lm et PCAlm.

3.7 Discussion

Les études d'association sont largement utilisées en biologie pour trouver des liens de corrélation entre des variables d'intérêt. Nous nous sommes intéressés à la situation où des variables non observées expliquent une partie des variations des variables expliquées et sont corrélées aux variables explicatives. Nous avons proposé dans ce chapitre deux algorithmes, ridgeLFMM et lassoLFMM, pour corriger les études d'association à l'aide de facteurs latents. Nos méthodes reposent sur les solutions optimales de problèmes de moindres carrés régularisés. La méthode ridgeLFMM utilise une régularisation L_2 portant sur la matrice des effets. La régularisation L_2 permet de définir de façon unique la matrice des variables latentes. La méthode lassoLFMM utilise une régularisation L_1 portant sur la matrice des effets. La régularisation L_1 permet de choisir a priori la proportion de variables expliquées associées à la variable d'intérêt. Les deux algorithmes d'estimation des facteurs de confusion que nous avons présentés reposent sur les deux mêmes étapes : on transforme la matrice des variables expliquées afin d'estimer les variables latentes, puis on estime les effets des variables explicatives sur les variables expliquées. L'algorithme ridgeLFMM fait une seule fois chaque étape, tandis que l'algorithme lassoLFMM les itère plusieurs fois.

Afin dévaluer l'intérêt d'estimer les facteurs latents dans les études d'association, nous avons considéré la méthode lm qui repose sur un modèle de régression de chaque variable expliquée par la variable explicative. Nous avons observé que les p-valeurs retournées par lm sont mal calibrées en présence de facteurs de confusion. En outre, la méthode lm a montré moins de puissance pour retrouver les candidats sur les données simulées et réelles. Cela montre que la présence de facteurs de confusion influence la calibration des p-valeurs ainsi que leurs rangs.

Nous avons considéré une méthode de correction pour les facteurs de confusion qui utilise les scores de l'ACP (PCAlm). Nous avons montré que la méthode PCAlm permet d'obtenir des p-valeurs correctement calibrées sur les simulations et les données réelles. Cependant nous avons observé sur les simulations que PCAlm retrouve moins bien les variables associées que les autres méthodes prenant en compte les variables

latentes. De même, sur les données réelles la méthodes PCAlm propose moins de candidats, passant ainsi à côté de certaines associations potentielles.

Les deux méthodes SVA comparées dans ce chapitre donnent des résultats similaires à PCAlm sur les simulations numériques. La méthode sva-two-step donne des résultats très similaires à PCAlm sur l'EWAS, alors que sva-two-step donne des résultats très atypiques comparés aux autres méthodes considérées pour l'EWAS. Avec les bons hyperparamètres, il est possible que sva-irw donne des résultats comparables aux autres méthodes. Cependant il n'y a aucune garantie sur la convergence de sva-irw ; ce qui rend compliqué l'exploration pour trouver les bons paramètres de l'algorithme.

Sur les jeux de données simulées, nos algorithmes ont les mêmes performances que la méthode oracle qui connait les facteurs de confusion. Les p-valeurs renvoyées par nos algorithmes sont bien calibrées sur les vrais jeux de données, et permettent de retrouver les candidats trouvés par d'autres études. Les performances globales de ces méthodes sont très comparables à celles de la méthode cate. Bien que cate, ridgeLFMM et lassoLFMM aient des performances très similaires pour nos critères d'évaluation, il existe tout de même des différences entre les listes de candidats renvoyées pas ces trois méthodes. Par ailleurs, il existe des techniques permettant de combiner les résultats venant de plusieurs méthodes pour tester l'association ; ce qui permettrait d'augmenter la confiance que l'on a en les résultats d'une étude d'association.

Dans les études d'association sur des données réelles, il n'y pas de vérité terrain. De plus, une méthode peut toujours donner de meilleurs résultats qu'une autre sur des simulations bien choisies (Wolpert et Macready, 1997). Sur les données réelles nous avons constaté que les méthodes cate, lassoLFMM et ridgeLFMM retournent des p-valeurs correctement calibrées qui permettent de calculer des listes pour des FDR controlé. Ces listes sont en général plus grandes que celles des autres méthodes. Nous avons implémenté nos algorithmes dans un package R. Ce qui permet aux utilisateurs de les combiner avec d'autres algorithmes d'association. Ces deux nouvelles méthodes s'ajoutent ainsi à l'arsenal des méthodes permettant de corriger les tests d'association statistique pour les facteurs de confusion.

ID	Chr	Position	Gene	PCAlm	lassoLFMM	cate	ridgeLFMM
cg16411857	**16**	**57023191**	**NLRC5**	**9.2e-13**	**2.4e-12**	**6.6e-12**	**5.3e-12**
cg07839457	**16**	**57023022**	**NLRC5**	**1.9e-11**	**4.5e-11**	**1.1e-10**	**9.7e-11**
cg05428452	**6**	**32712979**	**HLA-DQA2**	**5.4e-11**	**4.6e-11**	**8.5e-11**	**8.8e-11**
cg02508743	8	56903623	LYN	2.9e-08	2.7e-08	2.7e-08	2.8e-08
cg20821042	**6**	**32709158**	**HLA-DQA2**	**6.5e-08**	**6.1e-08**	**9.6e-08**	**1.0e-07**
cg13081526	6	32449961		1.5e-07	1.2e-07	2.0e-07	2.2e-07
cg18052547	6	32552547	HLA-DRB1	1.8e-07	1.8e-07	3.0e-07	3.1e-07
cg25372449	**6**	**32490350**	**HLA-DRB5**	**2.5e-07**	**2.6e-07**	**4.5e-07**	**4.6e-07**
cg02030958	13	110386267		4.0e-07	7.8e-08	6.0e-08	1.1e-07
cg16171858	3	58472734		4.6e-07	1.6e-07	2.7e-08	3.8e-08
cg03280622	8	145023013	PLEC1	4.7e-07	5.0e-09	5.8e-09	3.8e-08
cg24150157	19	51891210	LIM2	6.2e-07	3.1e-07	1.6e-07	2.1e-07
cg26244575	12	76354015		6.9e-07	2.7e-09	5.0e-10	4.2e-09
cg05370853	6	32606634	HLA-DQA1	7.1e-07	3.0e-07	3.3e-07	4.4e-07
cg14989316	10	80757927	LOC283050	7.3e-07	6.1e-08	7.8e-08	2.1e-07
cg17360552	6	32725332	HLA-DQB2	8.1e-07	6.1e-07	1.1e-06	1.2e-06
cg01373248	3	18480297	SATB1	8.1e-07	1.4e-07	1.1e-07	2.5e-07
cg26164488	2	64440295		9.3e-07	3.5e-09	1.6e-09	1.4e-08
cg05874806	2	102350276	MAP4K4	1.1e-06	1.1e-06	4.7e-07	5.6e-07

TABLE 3.1 – **Sites de méthylation associés avec la maladie polyarthrite rhumatoïde (EWAS).** p-valeurs des candidats détectés par les méthodes PCAlm, lassoLFMM, ridgeLFMM et cate pour un taux de fausse découverte controlé à 1% pour l'EWAS. Les Sites de méthylation en gras sont les candidats de (ZOU et al., 2014 ; RAHMANI et al., 2016)

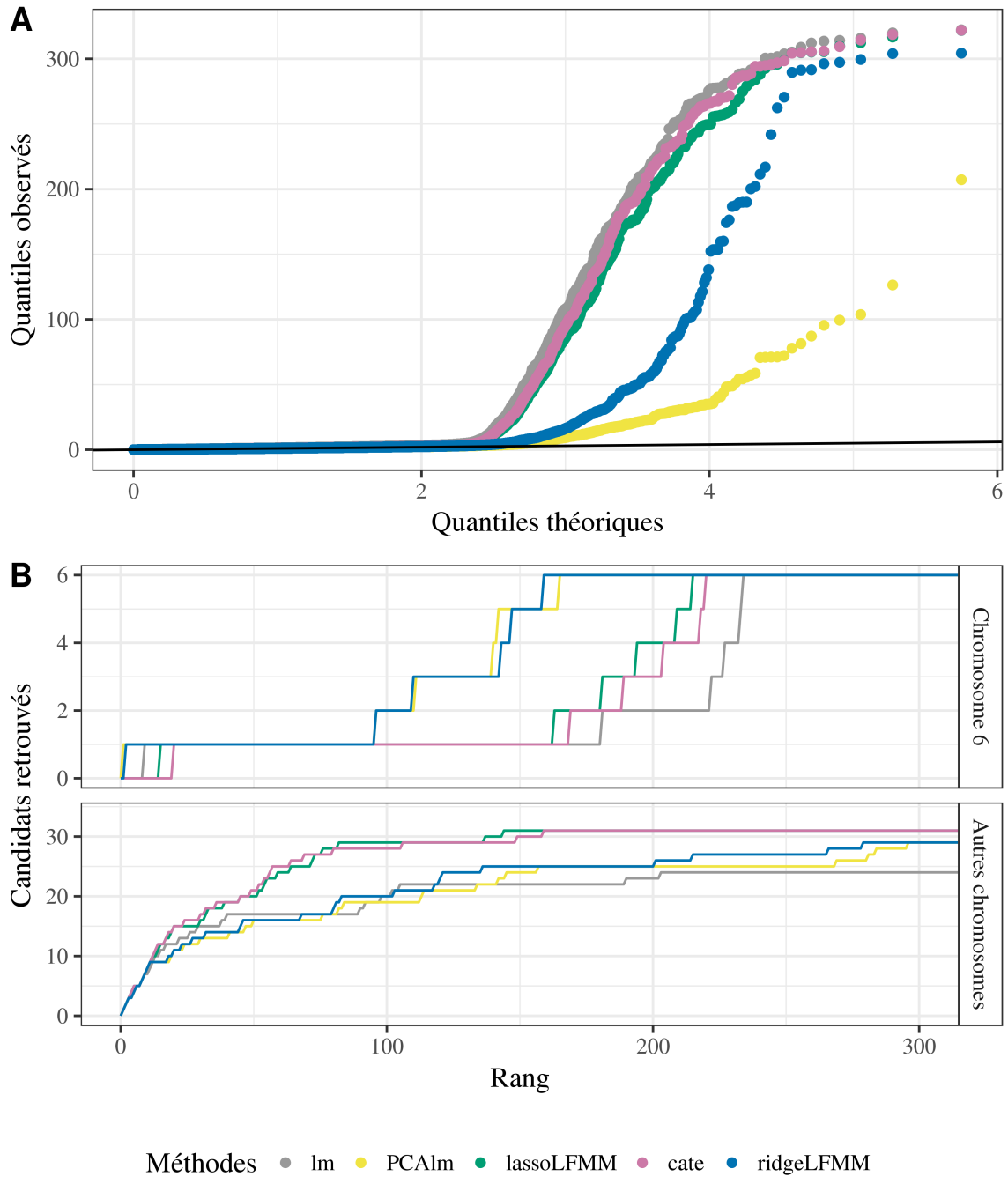

FIGURE 3.7 – **Q-Q plot et rang pour l'étude d'association entre des génomes et la maldie cœliaque (GWAS).** A) Diagramme quantile-quantile de l'inverse du logarithme en base 10 des p-valeurs renvoyées par chaque méthode. Les quantiles théoriques sont ceux de la loi exponentielle. B) Nombre de gènes du GWAS catalog retrouvés dans la liste rangée (la liste est rangée selon les p-valeurs) avant un rang donné.

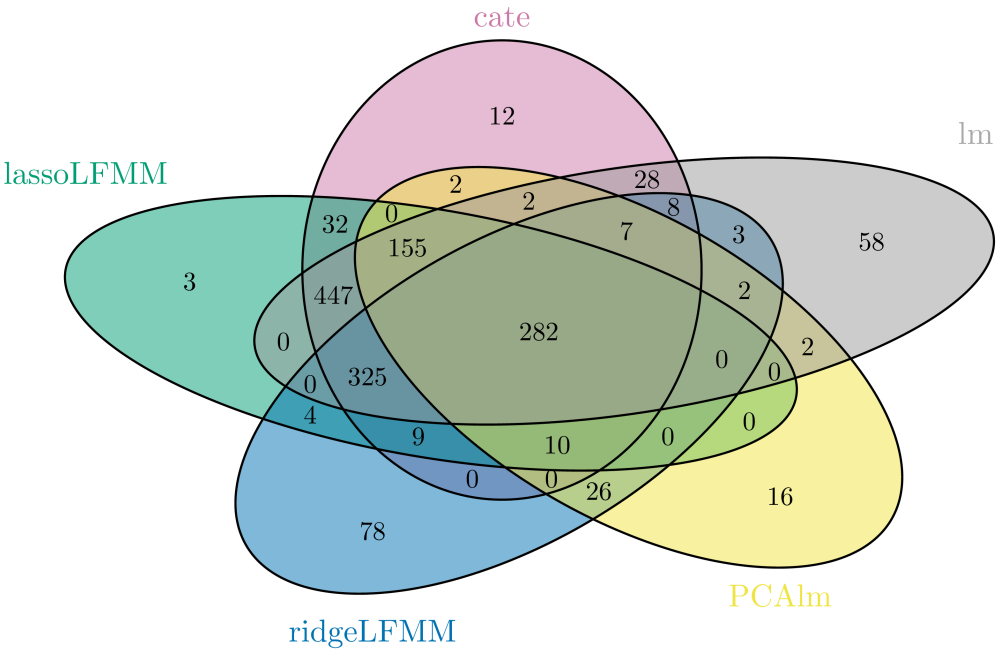

FIGURE 3.8 – **Diagramme de Venn de l'étude d'association entre des génotypes et la maladie cœliaque (GWAS).** Diagramme de Venn des listes controlées à un taux de fausses de découvertes de 1% pour chaque méthode.

Chr	Snp	LD block (Mb)	Odds ratio [95% IC]	p-valeur	q-valeur	Gènes dans la zone
3	**rs1464510**	**189.56-189.61**	**1.30 [1.24-1.36]**	**3.8e-23**	**1.5e-20**	**LPP**
3	**rs17810546**	**160.99-161.32**	**1.35 [1.26-1.45]**	**1.8e-16**	**6.1e-14**	**IQCJ-SCHIP1, IL12A-AS1, IL12A**
4	**rs13151961**	**123.19-123.56**	**0.73 [0.68-0.78]**	**1.7e-14**	**5.3e-12**	**KIAA1109, ADAD1**
12	**rs653178**	**110.25-110.49**	**1.22 [1.16-1.28]**	**6.8e-13**	**2.1e-10**	**CUX2, LINC02356, SH2B3, ATXN2**
2	**rs917997**	**102.26-102.61**	**1.27 [1.20-1.35]**	**1.5e-12**	**4.6e-10**	**IL1RL1, IL18R1, IL18RAP, MIR4772, SLC9A4, SLC9A2**
4	rs6840978	123.73-123.77	0.77 [0.72-0.82]	1.2e-11	3.5e-09	IL21-AS1
3	rs9811792	161.12-161.18	1.18 [1.12-1.24]	6.6e-11	1.9e-08	IL12A-AS1
3	**rs13098911**	**45.98-46.21**	**1.32 [1.22-1.43]**	**2.1e-10**	**5.8e-08**	**FYCO1, FLT1P1, CCR3**
1	**rs2816316**	**190.77-190.80**	**0.78 [0.72-0.83]**	**2.2e-10**	**6.2e-08**	
3	**rs6441961**	**46.26-46.33**	**1.21 [1.15-1.27]**	**1.7e-08**	**4.6e-06**	**CCR3, UQCRC2P1**
2	**rs4675374**	**204.29-204.52**	**1.23 [1.16-1.31]**	**2.1e-07**	**5.4e-05**	**CD28, ICOS**
2	rs1018326	181.54-181.78	1.15 [1.10-1.21]	4.4e-07	1.1e-04	UBE2E3, LINC01934
3	rs7648827	46.56-46.56	1.22 [1.12-1.33]	4.6e-07	1.2e-04	LRRC2
2	**rs13003464**	**60.95-61.24**	**1.19 [1.13-1.25]**	**5.0e-07**	**1.3e-04**	**LINC01185, REL, PUS10, RNA5SP95, KIAA1841, C2orf74**
10	**rs1250552**	**80.71-80.74**	**0.84 [0.80-0.88]**	**5.2e-07**	**1.3e-04**	**ZMIZ1**
3	rs7629708	189.56-189.62	1.17 [1.11-1.24]	1.0e-06	2.5e-04	LPP
22	**rs2298428**	**20.13-20.31**	**1.17 [1.10-1.24]**	**1.3e-06**	**3.3e-04**	**HIC2, UBE2L3, YDJC, CCDC116**
18	rs1394466	48.93-49.30	1.14 [1.08-1.20]	1.5e-06	3.6e-04	DCC
18	**rs1893217**	**12.80-12.84**	**1.16 [1.08-1.23]**	**1.6e-06**	**4.0e-04**	**PTPN2**
1	**rs864537**	**165.66-165.70**	**0.87 [0.83-0.92]**	**1.7e-06**	**4.2e-04**	**POU2F1, CD247**

TABLE 3.2 – **20 premiers groupes de SNPs associés avec la maladie retrounés par l'algorithme lassoLFMM.** Les lignes en gras sont les SNPs référencés dans la base de données GWAS catalog.

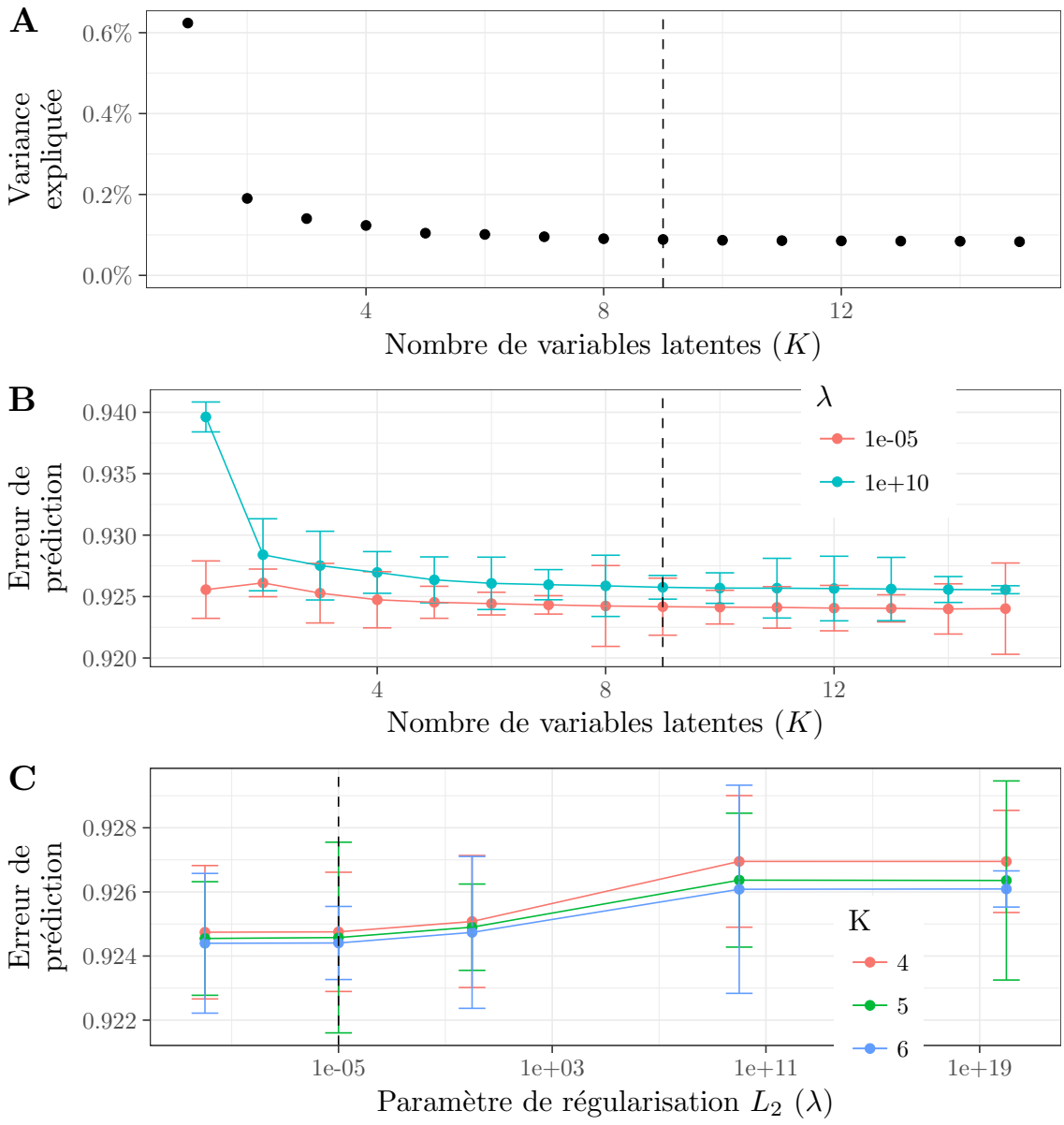

FIGURE 3.9 – **Choix des paramètres pour l'étude d'association entre génotype et un gradient environnemental (GEAS).** A) Proportion de variance expliquée de la projection de $\mathbf{Y}$ sur l'espace orthogonal à $\mathbf{X}$ (c'est à dire $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$) par chacune des composantes principales. B)C) Erreur de prédiction calculée grâce à la validation croisée des estimateurs L_2 des paramètres de LFMM pour différente valeurs du paramètre de régularisation λ et du nombre variable latentes K. Le point représente l'erreur de prédiction moyen et les barres l'erreur standard. La ligne pointillée verticale marque sur A et B le nombre de variables latentes choisies, c'est à dire 9, et sur C le paramètre de régularisation choisi, c'est à dire $\lambda = 10^{-3}$.

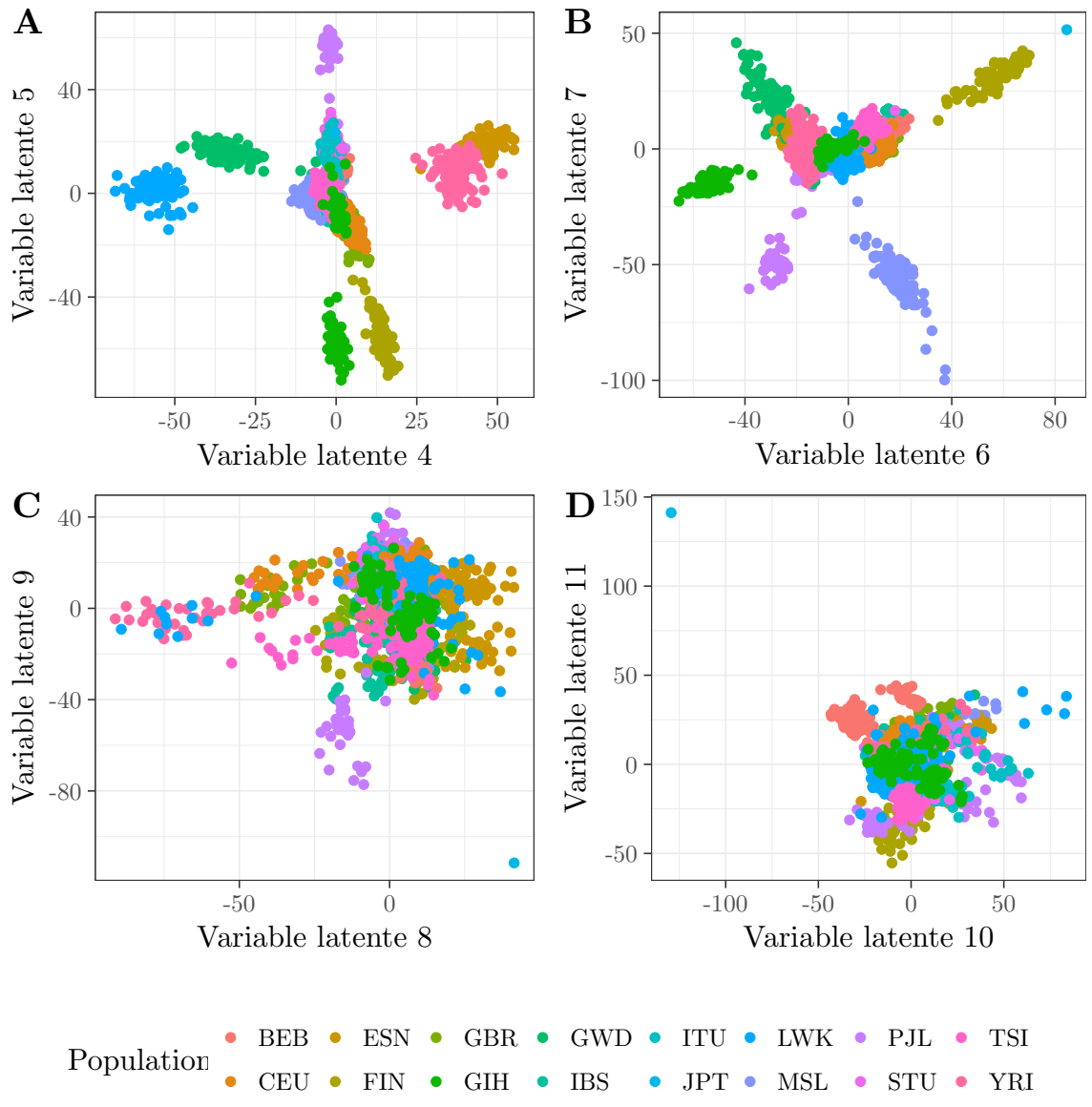

FIGURE 3.10 – **Variables latentes retournées par ridgeLFMM pour l'étude d'association entre génotype et un gradient environnemental (GEAS).** Les noms de populations sont les codes utilisés dans le projet 1000Genomes.

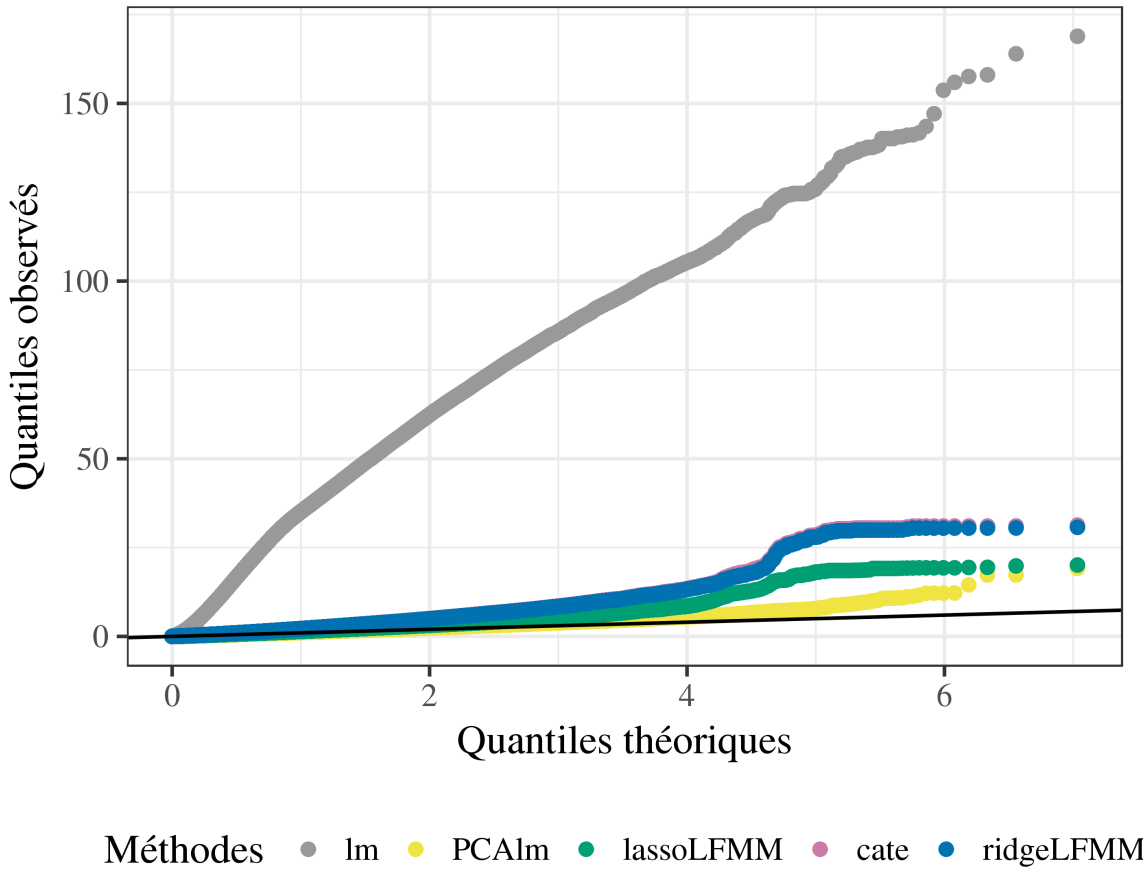

FIGURE 3.11 – **Q-Q plot de l'étude d'association entre génotype et un gradient environnemental (GEAS).** Diagramme quantile-quantile de l'inverse du logarithme en base 10 des p-valeurs renvoyées par chaques méthodes. Les quantiles théoriques suivent la loi exponentielle.

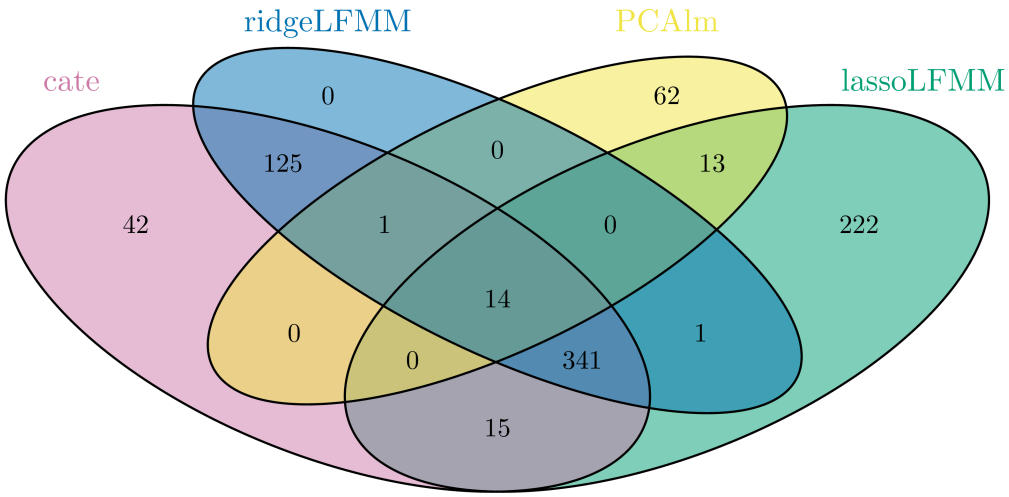

FIGURE 3.12 – **Diagramme de Venn de l'étude d'association entre des génotypes et un gradient environnemental (GEAS).** Diagramme de Venn des listes controlées à un taux de fausses de découvertes de 1% pour chaque méthode.

SNPs	Détecté par les méthodes	Description du phénotype
rs10908907	ridgeLFMM, cate	Alcoholism (heaviness of drinking)
rs10496731	lassoLFMM	Body Height
rs2472297	ridgeLFMM, cate, lassoLFMM	Caffeine metabolism
rs2256175	ridgeLFMM, cate, lassoLFMM	Cholesterol total
rs2472297	ridgeLFMM, cate, lassoLFMM	Coffee consumption (cups per day)
rs2278544, rs2322659	lassoLFMM	Congenital lactase deficiency
rs4954218	ridgeLFMM, cate, lassoLFMM	Corneal structure
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiographic traits
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiography
rs2256175	ridgeLFMM, cate, lassoLFMM	Giant cell arteritis
rs2256175, rs6085576, rs2104012, rs1983716, rs2853977	ridgeLFMM, cate, lassoLFMM	Height
rs6430549	ridgeLFMM, cate, lassoLFMM	Hematocrit
rs2278544, rs2322659	lassoLFMM	Lactose intolerance
rs882300	ridgeLFMM, cate, lassoLFMM	Multiple sclerosis
rs1123848	ridgeLFMM, cate, lassoLFMM	Neuroblastoma
rs17158483	lassoLFMM	Obesity-related traits

TABLE 3.3 – **SNPs associés avec un gradient climatique qui ont été associés avec des phénotypes dans d'autres études.** SNPs présents dans l'union des SNPs détectés par les méthodes lassoLFMM, ridgeLFMM, cate et PCAlm pour un le FDR à 1%.

Chapitre 4

Conclusions et perspectives

Un grande partie de la biologie repose sur l'analyse des données provenant du monde qui nous entoure. Les techniques qui permettent d'acquérir les données évoluent ; il est donc nécessaire que les méthodes pour les analyser évoluent également. Dans le cadre de cette thèse, nous avons proposé plusieurs algorithmes résolvant des problèmes de factorisation de matrice. Pour certains algorithmes nous avons prouvé leur convergence vers un point critique de leur fonction objectif. Nous avons démontré l'utilité de chacun de nos algorithmes sur des données réelles afin de répondre à des questions de la génétique moderne. Des implémentations efficaces de nos algorithmes sont disponibles dans deux packages R : `tess3r` et `lfmm`. Nous proposons maintenant différentes perspectives de nos travaux.

4.1 Développement et maintenance des packages `tess3r` et `lfmm`

Le choix de l'implémentation des packages en R n'est pas anodin. La communauté R est très vivante et en particulier en analyse de données biologiques. Les packages que nous avons développés peuvent ainsi être intégrés à des pipelines d'analyse de données biologiques faisant intervenir d'autre package R. Par ailleurs, le langage de programmation R est un langage "haut niveau" permettant une maintenance plus facile que sur des langages de programmation bas niveau tel que le langage C par exemple. Une perspective majeure de nos travaux est de maintenir nos packages en fonction des mises à jour du langage R, ou bien des rapports d'erreurs renvoyés par les utilisateurs. Un autre aspect important du développement des packages est la mise

en place d'une documentation complète. Au même titre que le développement des fonctions informatiques, la documentation doit évoluer avec les retours des utilisateurs.

Les packages `tess3r` et `lfmm` reposent sur des structures de données matricielles. Nous avons choisi d'utiliser la structure matricielle classique du langage R afin d'être le plus général possible. Une question majeure pour les utilisateurs est de savoir comment charger les données en mémoire vive afin d'utiliser nos outils. Certains formats de fichier se sont imposés dans la communauté de la génétique, comme par exemple le format `.bed` du logiciel `plink`, ou le format `.vcf`. Ajouter une option permettant aux utilisateurs de fournir directement un fichier formaté dans un format standard, serait un réel avantage pour les utilisateurs. De plus, certains formats de fichier ont été pensés pour permettre un accès rapide des données à la manière des algorithmes onlines [1]. C'est le cas par exemple du format `.bed`. Ne pas charger toutes les données dans la mémoire vive de l'ordinateur permettrait aux utilisateurs ayant des capacités en mémoire vive modeste d'utiliser nos algorithmes sur des données massives.

Enfin, pour rendre nos outils utilisables par le plus grand nombre, il est très important de communiquer sur ceux-ci. En particulier, il faut inciter les utilisateurs à poser des questions sur les plateformes d'échange du type Stack Exchange comme biostars et stackoverflow. Ces plateformes permettent des échanges avec les utilisateurs, à la fois sur les aspects pratiques d'utilisation des packages, mais aussi sur les aspects plus généralistes d'analyse de données. Ces échanges permettent de faire le lien entre statisticiens et biologistes et stimule la recherche de nouvelles méthodes statistiques.

4.2 Valeurs aberrantes et données manquantes

Les données réelles analysées lors de cette thèse peuvent être qualifiées de données "propres" ; c'est-à-dire que les données contiennent peu d'erreurs et peu de valeurs manquantes. Cependant, il est fréquent sur les grands jeux de données biologiques avec beaucoup de variables, qu'un certain nombre de variables ne soit pas observé pour tous les échantillons. Ou bien qu'il y ait des erreurs dans les variables observées. Par exemple, il est fréquent que les données génétiques comportent des erreurs de

1. Un algorithme online, ou algorithme en ligne, est un algorithme qui reçoit son entrée non pas d'un seul coup, mais comme un flux de données, et qui doit prendre des décisions au fur et à mesure.

séquençage. Des données génétiques de très mauvaise qualité sont par exemple les données génétiques prélevées sur des espèces qui se sont éteintes (ADN fossile).

Les algorithmes que nous avons développés ont été pensés pour fonctionner sur des matrices de données complètes. Dans le cas des données génétiques, l'imputation des valeurs manquantes a été très étudiée (B. L. Browning et S. R. Browning, 2016). L'approche la plus simple est d'imputer les valeurs manquantes avant de considérer d'autres analyses (d'association ou d'inférence de la structure de population par exemple). Cette approche fonctionne très bien pour les espèces très étudiées comme l'humain. Pour de telles espèces, on a accès à beaucoup de génomes complets pouvant servir de référence pour l'imputation (Howie et al., 2012). Pour des espèces anciennes ou simplement moins étudiées l'imputation est moins évidente.

Les algorithmes que nous avons développés peuvent être découpés en une suite d'opérations matricielles. Les opérations matricielles n'étant pas définies avec des matrices incomplètes, il est nécessaire d'adapter les algorithmes pour qu'ils gèrent les valeurs manquantes. Nos algorithmes reposent sur des fonctions objectif. Il est donc possible d'écrire les fonctions objectif avec des matrices incomplètes. Nous perdons cependant les structures matricielles des variables lors de l'optimisation des fonctions objectif. Il faut alors envisager d'autres algorithmes d'optimisation que les algorithmes de descente par blocs de coordonnées utilisés dans nos méthodes. Une autre approche simple à implémenter consiste à imputer les variables manquantes par leur moyenne ou une autre statistique. Quand il y a peu de données manquantes, on s'attend à ce que cette méthode ne biaise pas trop l'inférence des facteurs dans les méthodes à facteurs comme l'ACP (Josse et al., 2009).

Une autre approche consiste à imputer au hasard une première fois les données manquantes. Puis dans le cadre d'un algorithme d'estimation itératif à chaque étape on utilise la valeur prédite par le modèle. Il s'agit de l'approche utilisée par l'algorithme SOFT-IMPUTE prévu pour fonctionner sur des matrices de données très incomplètes (Mazumder et al., 2010). Cette approche pourrait être directement adaptable aux algorithmes APLS, AQP et lassoLFMM, ceux-ci étant des algorithmes itératifs. Nous perdons cependant les propriétés de convergence démontrées sur les algorithmes AQP et lassoLFMM.

On attribue parfois un score à la confiance des séquençages génétiques. Une méthode simple consiste à filtrer les observations ayant un score trop faible pour les traiter

comme des données manquantes. Le problème est autrement plus complexe lorsque que nous ne sommes pas capables de détecter les erreurs de mesure présentes dans les données biologiques. Nous avons évalué, dans le chapitre 2, l'effet du bruit dans les données géographiques sur les résultats de nos algorithmes spatiaux d'estimation de l'ascendance. Cependant, nous pouvons nous interroger sur l'effet des valeurs aberrantes présentes dans les données génétiques, ainsi que sur la façon dont les erreurs de mesure peuvent influencer les études d'association faites à partir de notre package `lfmm`.

4.3 Conclusion générale

Dans cette thèse, nous avons développé deux logiciels d'analyse de données répondant à des problèmes spécifiques de la biologie. Nos travaux s'inscrivent dans une logique de coévolution des méthodes statistiques et des techniques d'acquisition de données. Les perspectives ouvertes par cette thèse sont multiples. D'une part, nos travaux ouvrent des perspectives de développement méthodologique et logiciel, afin par exemple d'ajouter des fonctionnalités à nos packages R en fonction des retours des utilisateurs. D'autre part, nos travaux ouvrent des perspectives sur l'utilisation de la factorisation matricielle dans d'autres problèmes d'analyse de données massives. Nous pensons par exemple aux études d'association à partir de données d'expression génique (données RNA-Seq), ou bien de données de méthylation de l'ADN utilisant les technologies de séquençage dernière génération.

> Arrakis enseigne l'attitude du couteau : couper ce qui est incomplet et dire : "Maintenant, c'est complet, car cela s'achève ici."
>
> princesse Irulan

Bibliographie

Agarwal, Deepak et Bee-Chung Chen (- 2009). "Regression-based latent factor models". In : *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, nil. DOI : 10.1145/1557019.1557029. URL : https://doi.org/10.1145/1557019.1557029 (cf. p. 12, 13, 53).

Alexander, D. H. et Kenneth Lange (2011). "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation". In : *BMC Bioinformatics* 12.1, p. 246. ISSN : 1471-2105. DOI : 10.1186/1471-2105-12-246. URL : http://dx.doi.org/10.1186/1471-2105-12-246 (cf. p. 8, 21, 30).

Alexander, D. H., John Novembre et K. Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". In : *Genome Research* 19, p. 1655-1664. DOI : 10.1101/gr.094052.109 (cf. p. 7, 8).

Balding, David J. (oct. 2006). "A tutorial on statistical methods for population association studies". In : *Nature Reviews Genetics* 7.10, p. 781-791. ISSN : 1471-0064. DOI : 10.1038/nrg1916. URL : http://dx.doi.org/10.1038/nrg1916 (cf. p. 50).

Baran, Yael et al. (juin 2013). "Enhanced Localization of Genetic Samples through Linkage-Disequilibrium Correction". In : *The American Journal of Human Genetics* 92.6, p. 882-894. ISSN : 0002-9297. DOI : 10.1016/j.ajhg.2013.04.023. URL : http://dx.doi.org/10.1016/j.ajhg.2013.04.023 (cf. p. 22).

Benjamini, Yoav et Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In : *Journal of the royal statistical society. Series B (Methodological)*, p. 289-300 (cf. p. 31, 73).

Bertsekas, D P (mar. 1997). "Nonlinear Programming". In : *Journal of the Operational Research Society* 48.3, p. 334-334. ISSN : 1476-9360. DOI : 10.1057/palgrave.jors.2600425. URL : http://dx.doi.org/10.1057/palgrave.jors.2600425 (cf. p. 22, 26, 60).

Bhaskar, Anand et al. (déc. 2016). "Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies". In : *Bioinformatics*, btw720. ISSN : 1460-2059. DOI : 10.1093/bioinformatics/btw720. URL : http://dx.doi.org/10.1093/bioinformatics/btw720 (cf. p. 22).

Bradburd, Gideon, Graham Coop et Peter Ralph (2017). "Inferring Continuous and Discrete Population Genetic Structure Across Space". In : *bioRxiv*. DOI : 10.1101/189688. eprint : http://www.biorxiv.org/content/early/2017/09/15/189688.full.pdf. URL : http://www.biorxiv.org/content/early/2017/09/15/189688 (cf. p. 8).

Bro, R. et al. (jan. 2008). "Cross-validation of component models: A critical look at current methods". In : *Analytical and Bioanalytical Chemistry* 390.5, p. 1241-1251. ISSN : 1618-2650. DOI : 10.1007/s00216-007-1790-1. URL : http://dx.doi.org/10.1007/s00216-007-1790-1 (cf. p. 66).

Browning, Brian L. et Sharon R. Browning (jan. 2016). "Genotype Imputation with Millions of Reference Samples". In : *The American Journal of Human Genetics* 98.1, p. 116-126. ISSN : 0002-9297. DOI : 10.1016/j.ajhg.2015.11.020. URL : http://dx.doi.org/10.1016/j.ajhg.2015.11.020 (cf. p. 76, 101).

Cai, Deng et al. (août 2011). "Graph Regularized Nonnegative Matrix Factorization for Data Representation". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, p. 1548-1560. ISSN : 0162-8828. DOI : 10.1109/TPAMI.2010.231. URL : http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5674058 (cf. p. 25).

Cai, Jian-Feng, Emmanuel J. Candès et Zuowei Shen (2010). "A Singular Value Thresholding Algorithm for Matrix Completion". In : *SIAM Journal on Optimization* 20.4, p. 1956-1982. DOI : 10.1137/080738970. URL : https://doi.org/10.1137/080738970 (cf. p. 60).

Carbon, Seth et al. (nov. 2008). "AmiGO: online access to ontology and annotation data". In : *Bioinformatics* 25.2, p. 288-289. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btn615. URL : http://dx.doi.org/10.1093/bioinformatics/btn615 (cf. p. 34).

Carvalho, Carlos M. et al. (2008). "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics". In : *Journal of the American Statistical Association* 103.484, p. 1438-1456. DOI : 10.1198/016214508000000869. URL : https://doi.org/10.1198/016214508000000869 (cf. p. 52).

Cavalli, Luigi Luca, Paolo Menozzi et Alberto Piazza (1994). *The History and Geography of Human Genes*. Princeton University Press, p. 1059. ISBN : 9780691087504 (cf. p. 43).

Chen, Chibiao et al. (sept. 2007). "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study". In : *Molecular Ecology Notes* 7.5, p. 747-756. ISSN : 1471-8286. DOI : 10.1111/j.1471-8286.2007.01769.x. URL : http://dx.doi.org/10.1111/j.1471-8286.2007.01769.x (cf. p. 7, 8, 16, 22).

Choi, Yunjin, Jonathan Taylor et Robert Tibshirani (2014). "Selecting the number of principal components: Estimation of the true rank of a noisy matrix". In : *arXiv preprint arXiv:1410.8260* (cf. p. 64).

Cichocki, Andrzej et al. (sept. 2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Chichester, UK : John Wiley & Sons, Ltd, p. 1-477. ISBN : 9780470746660. DOI : 10.1002/9780470747278. arXiv : MSP.2007.911394 [10.1109]. URL : http://doi.wiley.com/10.1002/9780470747278 (cf. p. 24).

Consortium, The 1000 Genomes Project (sept. 2015). "A global reference for human genetic variation". In : *Nature* 526.7571, p. 68-74. DOI : 10.1038/nature15393. URL : https://doi.org/10.1038%2Fnature15393 (cf. p. 73, 76).

Corander, Jukka, Jukka Sirén et Elja Arjas (jan. 2008). "Bayesian spatial modeling of genetic population structure". In : *Computational Statistics* 23.1, p. 111-129. issn : 0943-4062. doi : 10.1007/s00180-007-0072-x. url : http://link.springer.com/10.1007/s00180-007-0072-x (cf. p. 7, 8, 22).

Cressie, Noel A. C. (sept. 1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA : John Wiley & Sons, Inc. isbn : 9781119115151. doi : 10.1002/9781119115151. url : http://doi.wiley.com/10.1002/9781119115151 (cf. p. 29).

Darwin, Charles (1859). "On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life." In : doi : 10.5962/bhl.title.2109. url : http://dx.doi.org/10.5962/bhl.title.2109 (cf. p. 76).

Devlin, B. et Kathryn Roeder (déc. 1999). "Genomic Control for Association Studies". In : *Biometrics* 55.4, p. 997-1004. issn : 0006-341X. doi : 10.1111/j.0006-341x.1999.00997.x. url : http://dx.doi.org/10.1111/j.0006-341x.1999.00997.x (cf. p. 31, 73).

Dubois, Patrick CA et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression". In : *Nature genetics* 42.4, p. 295-302. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847618/ (cf. p. 75, 84).

Durand, E. et al. (mai 2009). "Spatial Inference of Admixture Proportions and Secondary Contact Zones". In : *Molecular Biology and Evolution* 26.9, p. 1963-1973. issn : 1537-1719. doi : 10.1093/molbev/msp106. url : http://dx.doi.org/10.1093/molbev/msp106 (cf. p. 14, 22, 32).

Eastment, H. T. et W. J. Krzanowski (fév. 1982). "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis". In : *Technometrics* 24.1, p. 73-77. issn : 1537-2723. doi : 10.1080/00401706.1982.10487712. url : http://dx.doi.org/10.1080/00401706.1982.10487712 (cf. p. 30).

Eckart, Carl et Gale Young (sept. 1936). "The approximation of one matrix by another of lower rank". In : *Psychometrika* 1.3, p. 211-218. issn : 1860-0980. doi : 10.1007/bf02288367. url : http://dx.doi.org/10.1007/bf02288367 (cf. p. 57).

Efron, Bradley (mar. 2004). "Large-Scale Simultaneous Hypothesis Testing". In : *Journal of the American Statistical Association* 99.465, p. 96-104. issn : 1537-274X. doi : 10.1198/016214504000000089. url : http://dx.doi.org/10.1198/016214504000000089 (cf. p. 68).

Engelhardt, Barbara E. et Matthew Stephens (sept. 2010). "Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis". In : *PLoS Genetics* 6.9. Sous la dir. de BruceEditor Walsh, e1001117. issn : 1553-7404. doi : 10.1371/journal.pgen.1001117. url : http://dx.doi.org/10.1371/journal.pgen.1001117 (cf. p. 21).

Epperson, Bryan K. et Tianquan Li (sept. 1996). "Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics." In : *Proceedings of the National Academy of Sciences* 93.19, p. 10528-10532. issn : 1091-6490. doi : 10.1073/pnas.93.19.10528. url : http://dx.doi.org/10.1073/pnas.93.19.10528 (cf. p. 29, 43).

FALUSH, Daniel, Matthew STEPHENS et Jonathan K . PRITCHARD (2003). "Inference of Population Structure Using Multilocus Genotype Dat a: Linked Loci and Correlated Allele Frequencies". In : *Genetics* 164.4, p. 1567-1587. ISSN : 0016-6731. eprint : http://www.genetics.org/content/164/4/1567.full.pdf. URL : http://www.genetics.org/content/164/4/1567 (cf. p. 8).

FICK, Stephen E. et Robert J. HIJMANS (mai 2017). "Worldclim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas". In : *International Journal of Climatology*. ISSN : 0899-8418. DOI : 10.1002/joc.5086. URL : https://doi.org/10.1002/joc.5086 (cf. p. 76).

FISHER, Ronald Aylmer (1937). *The design of experiments*. Oliver et Boyd; Edinburgh; London (cf. p. 9).

FOURNIER-LEVEL, A et al. (oct. 2011). "A map of local adaptation in Arabidopsis thaliana". In : *Science (New York, N.Y.)* 334.6052, p. 86-89. ISSN : 1095-9203. DOI : 10.1126/science.1209271. URL : http://www.ncbi.nlm.nih.gov/pubmed/21980109 (cf. p. 41).

FRANÇOIS, OLIVIER et ERIC DURAND (août 2010). "Spatially explicit Bayesian clustering models in population genetics". In : *Molecular Ecology Resources* 10.5, p. 773-784. ISSN : 1755-098X. DOI : 10.1111/j.1755-0998.2010.02868.x. URL : http://dx.doi.org/10.1111/j.1755-0998.2010.02868.x (cf. p. 22, 32).

FRANÇOIS, Olivier, Helena MARTINS et al. (jan. 2016). "Controlling false discoveries in genome scans for selection". In : *Molecular Ecology* 25.2, p. 454-469. ISSN : 0962-1083. DOI : 10.1111/mec.13513. URL : http://dx.doi.org/10.1111/mec.13513 (cf. p. 31).

FRANÇOIS, Olivier et Lisette P. WAITS (sept. 2015). "Clustering and Assignment Methods in Landscape Genetics". In : *Landscape Genetics*, p. 114-128. DOI : 10.1002/9781118525258.ch07. URL : http://dx.doi.org/10.1002/9781118525258.ch07 (cf. p. 20).

FRICHOT, Eric et Olivier FRANÇOIS (mai 2015). "LEA: AnRpackage for landscape and ecological association studies". In : *Methods in Ecology and Evolution* 6.8. Sous la dir. de BrianEditor O'MEARA, p. 925-929. ISSN : 2041-210X. DOI : 10.1111/2041-210x.12382. URL : http://dx.doi.org/10.1111/2041-210x.12382 (cf. p. 6, 14, 34).

FRICHOT, Eric, François MATHIEU et al. (fév. 2014). "Fast and Efficient Estimation of Individual Ancestry Coefficients". In : *Genetics* 196.4, p. 973-983. ISSN : 1943-2631. DOI : 10.1534/genetics.113.160572. URL : http://dx.doi.org/10.1534/genetics.113.160572 (cf. p. 7, 8, 16, 21, 23, 24, 30, 34, 38).

FRICHOT, Eric, Sean D. SCHOVILLE et al. (mar. 2013). "Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models". In : *Molecular Biology and Evolution* 30.7, p. 1687-1699. ISSN : 0737-4038. DOI : 10.1093/molbev/mst063. URL : http://dx.doi.org/10.1093/molbev/mst063 (cf. p. 12, 13, 52).

FRIEDMAN, Jerome, Trevor HASTIE, Holger HÖFLING et al. (déc. 2007). "Pathwise coordinate optimization". In : *The Annals of Applied Statistics* 1.2, p. 302-332. ISSN : 1932-6157. DOI : 10.1214/07-aoas131. URL : http://dx.doi.org/10.1214/07-aoas131 (cf. p. 59).

FRIEDMAN, Jerome, Trevor HASTIE et Robert TIBSHIRANI (2010). "Regularization Paths for Generalized Linear Models Via Coordinate Descent". In : *Journal of Statistical Software* 33.1, nil. DOI : 10.18637/jss.v033.i01. URL : https://doi.org/10.18637/jss.v033.i01 (cf. p. 65).

FRIGUET, Chloé, Maela KLOAREG et David CAUSEUR (2009). "A Factor Model Approach To Multiple Testing Under Dependence". In : *Journal of the American Statistical Association* 104.488, p. 1406-1415. DOI : 10.1198/jasa.2009.tm08332. URL : https://doi.org/10.1198/jasa.2009.tm08332 (cf. p. 13, 53).

GERARD, David et Matthew STEPHENS (2017a). "Empirical Bayes Shrinkage and False Discovery Rate Estimation, Allowing For Unwanted Variation". In : *arXiv preprint arXiv:1709.10066* (cf. p. 12, 13).

— (2017b). "Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls". In : *arXiv preprint arXiv:1705.08393* (cf. p. 53, 67).

GRIPPO, L. et M. SCIANDRONE (avr. 2000). "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints". In : *Operations Research Letters* 26.3, p. 127-136. ISSN : 0167-6377. DOI : 10.1016/s0167-6377(99)00074-7. URL : http://dx.doi.org/10.1016/s0167-6377(99)00074-7 (cf. p. 15, 27).

GUEDJ, Benjamin et Gilles GUILLOT (juil. 2011). "Estimating the location and shape of hybrid zones". In : *Molecular Ecology Resources* 11.6, p. 1119-1123. ISSN : 1755-098X. DOI : 10.1111/j.1755-0998.2011.03045.x. URL : http://dx.doi.org/10.1111/j.1755-0998.2011.03045.x (cf. p. 7, 8).

GUJRAL, Naiyana (2012). "Celiac disease: Prevalence, diagnosis, pathogenesis and treatment". In : *World Journal of Gastroenterology* 18.42, p. 6036. ISSN : 1007-9327. DOI : 10.3748/wjg.v18.i42.6036. URL : http://dx.doi.org/10.3748/wjg.v18.i42.6036 (cf. p. 75).

HALKO, N., P. G. MARTINSSON et J. A. TROPP (jan. 2011). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In : *SIAM Review* 53.2, p. 217-288. ISSN : 1095-7200. DOI : 10.1137/090771806. URL : http://dx.doi.org/10.1137/090771806 (cf. p. 62).

HANCOCK, Angela M et al. (oct. 2011). "Adaptation to climate across the Arabidopsis thaliana genome." In : *Science (New York, N.Y.)* 334.6052, p. 83-86. ISSN : 1095-9203. DOI : 10.1126/science.1209244. URL : http://www.ncbi.nlm.nih.gov/pubmed/21980108 (cf. p. 41).

HARDY, Olivier J. (août 1999). "Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models". In : *Heredity* 83.2, p. 145. ISSN : 1365-2540. DOI : 10.1046/j.1365-2540.1999.00558.x. URL : http://dx.doi.org/10.1046/j.1365-2540.1999.00558.x (cf. p. 29).

HASTIE, Trevor, Robert TIBSHIRANI et Jerome FRIEDMAN (2009). "The Elements of Statistical Learning". In : *Springer Series in Statistics*. ISSN : 2197-568X. DOI : 10.1007/978-0-387-84858-7. URL : http://dx.doi.org/10.1007/978-0-387-84858-7 (cf. p. 68).

HORTON, Matthew W et al. (jan. 2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel". In : *Nature*

Genetics 44.2, p. 212-216. ISSN : 1546-1718. DOI : 10.1038/ng.1042. URL : http://dx.doi.org/10.1038/ng.1042 (cf. p. 16, 23, 34, 42, 44).

HOWIE, Bryan et al. (juil. 2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". In : *Nature Genetics* 44.8, p. 955-959. ISSN : 1546-1718. DOI : 10.1038/ng.2354. URL : http://dx.doi.org/10.1038/ng.2354 (cf. p. 101).

HUDSON, R. R. (fév. 2002). "Generating samples under a Wright-Fisher neutral model of genetic variation". In : *Bioinformatics* 18.2, p. 337-338. ISSN : 1460-2059. DOI : 10.1093/bioinformatics/18.2.337. URL : http://dx.doi.org/10.1093/bioinformatics/18.2.337 (cf. p. 32).

JOLLIFFE, Ian T (1986). "Principal Component Analysis". In : *Principal component analysis*. Springer, p. 115-128 (cf. p. 64).

JOSSE, Julie, J PAGES et F HUSSON (2009). "Gestion des données manquantes en analyse en composantes principales". In : *Journal de la Société Française de Statistique* 150.2, p. 28-51 (cf. p. 101).

KANG, H. M. et al. (fév. 2008). "Efficient Control of Population Structure in Model Organism Association Mapping". In : *Genetics* 178.3, p. 1709-1723. ISSN : 0016-6731. DOI : 10.1534/genetics.107.080101. URL : http://dx.doi.org/10.1534/genetics.107.080101 (cf. p. 12, 52).

LEE, Daniel D et H Sebastian SEUNG (1999). "Learning the parts of objects by non-negative matrix factorization". In : *Nature* 401.6755, p. 788. URL : https://www.ncbi.nlm.nih.gov/pubmed/10548103 (cf. p. 24).

LEEK, J. T. et J. D. STOREY (2007). "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis". In : *PLoS Genetics* 3.9, e161. ISSN : 1553-7404. DOI : 10.1371/journal.pgen.0030161. URL : http://dx.doi.org/10.1371/journal.pgen.0030161 (cf. p. 12, 13, 53, 69).

— (nov. 2008). "A general framework for multiple testing dependence". In : *Proceedings of the National Academy of Sciences* 105.48, p. 18718-18723. ISSN : 1091-6490. DOI : 10.1073/pnas.0808709105. URL : http://dx.doi.org/10.1073/pnas.0808709105 (cf. p. 13, 67, 69).

LEWONTIN, R. C. et Jesse KRAKAUER (1973). "DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE S ELECTIVE NEUTRALITY OF POLYMORPHISMS". In : *Genetics* 74.1, p. 175-195. ISSN : 0016-6731. eprint : http://www.genetics.org/content/74/1/175.full.pdf. URL : http://www.genetics.org/content/74/1/175 (cf. p. 30).

LI, J. Z. et al. (fév. 2008). "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation". In : *Science* 319.5866, p. 1100-1104. ISSN : 1095-9203. DOI : 10.1126/science.1153717. URL : http://dx.doi.org/10.1126/science.1153717 (cf. p. 20).

LIU, Yun et al. (jan. 2013). "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis". In : *Nature Biotechnology* 31.2, p. 142-147. ISSN : 1546-1696. DOI : 10.1038/nbt.2487. URL : http://dx.doi.org/10.1038/nbt.2487 (cf. p. 74).

LOH, Po-Ru et al. (2017). "Mixed model association for biobank-scale data sets". In : *bioRxiv*. DOI : 10.1101/194944. eprint : https://www.biorxiv.org/content/

early/2017/09/27/194944.full.pdf. URL : https://www.biorxiv.org/content/early/2017/09/27/194944 (cf. p. 12).

MacArthur, Jacqueline et al. (nov. 2016). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In : *Nucleic Acids Research* 45.D1, p. D896-D901. ISSN : 1362-4962. DOI : 10.1093/nar/gkw1133. URL : http://dx.doi.org/10.1093/nar/gkw1133 (cf. p. 75).

Malécot, Gustave (1948). *Les mathématiques de l'hérédité*. Paris : Masson et Cie (cf. p. 43).

Mantel, Nathan (1967). "The detection of disease clustering and a generalized regression approach". In : *Cancer research* 27.2 Part 1, p. 209-220 (cf. p. 29).

Marchini, Jonathan et al. (2004). "The effects of human population structure on large genetic association studies". In : *Nature genetics* 36.5, p. 512. URL : http://www.nature.com/ng/journal/v36/n5/full/ng1337.html (cf. p. 20).

Martins, Helena et al. (sept. 2016). "Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics". In : *Molecular Ecology* 25.20, p. 5029-5042. ISSN : 0962-1083. DOI : 10.1111/mec.13822. URL : http://dx.doi.org/10.1111/mec.13822 (cf. p. 16, 31, 34, 44).

Mazumder, Rahul, Trevor Hastie et Robert Tibshirani (2010). "Spectral Regularization Algorithms for Learning Large Incomplete Matrices". In : *Journal of machine learning research* 11.Aug, p. 2287-2322 (cf. p. 101).

McLaren, William et al. (mar. 2016). "The Ensembl Variant Effect Predictor". In : DOI : 10.1101/042374. URL : http://dx.doi.org/10.1101/042374 (cf. p. 77).

Mishra, B. et al. (jan. 2013). "Low-Rank Optimization with Trace Norm Penalty". In : *SIAM Journal on Optimization* 23.4, p. 2124-2149. ISSN : 1095-7189. DOI : 10.1137/110859646. URL : http://dx.doi.org/10.1137/110859646 (cf. p. 58).

Novembre, John et al. (nov. 2008). "Genes mirror geography within Europe". In : *Nature* 456.7219, p. 274-274. ISSN : 1476-4687. DOI : 10.1038/nature07566. URL : http://dx.doi.org/10.1038/nature07566 (cf. p. 21).

Patterson, Nick, Alkes L. Price et David Reich (2006). "Population Structure and Eigenanalysis". In : *PLoS Genetics* 2.12, e190. ISSN : 1553-7404. DOI : 10.1371/journal.pgen.0020190. URL : http://dx.doi.org/10.1371/journal.pgen.0020190 (cf. p. 4).

Popescu, Andrei-Alin et al. (oct. 2014). "A Novel and Fast Approach for Population Structure Inference Using Kernel-PCA and Optimization". In : *Genetics* 198.4, p. 1421-1431. ISSN : 1943-2631. DOI : 10.1534/genetics.114.171314. URL : http://dx.doi.org/10.1534/genetics.114.171314 (cf. p. 7, 8, 21).

Price, Alkes L et al. (juil. 2006). "Principal components analysis corrects for stratification in genome-wide association studies". In : *Nature Genetics* 38.8, p. 904-909. ISSN : 1061-4036. DOI : 10.1038/ng1847. URL : http://dx.doi.org/10.1038/ng1847 (cf. p. 12, 51, 67, 69, 75).

Pritchard, Jonathan K, Matthew Stephens et Peter Donnelly (juin 2000). "Inference of population structure using multilocus genotype data". In : *Genetics* 155.2, p. 945-959. ISSN : 00166731. DOI : 10.1111/j.1471-8286.2007.01758.x. URL : http://www.ncbi.nlm.nih.gov/pubmed/10835412%20http://www.

pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1461096 (cf. p. 5, 7, 8, 20).

Purcell, Shaun et al. (sept. 2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In : *The American Journal of Human Genetics* 81.3, p. 559-575. ISSN : 0002-9297. DOI : 10.1086/519795. URL : http://dx.doi.org/10.1086/519795 (cf. p. 75).

Rahmani, Elior et al. (mar. 2016). "Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies". In : *Nature Methods* 13.5, p. 443-445. ISSN : 1548-7105. DOI : 10.1038/nmeth.3809. URL : http://dx.doi.org/10.1038/nmeth.3809 (cf. p. 12, 51, 67, 74, 79, 82, 83, 89).

Raj, Anil, Matthew Stephens et Jonathan K. Pritchard (avr. 2014). "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". In : *Genetics* 197.2, p. 573-589. ISSN : 1943-2631. DOI : 10.1534/genetics.114.164350. URL : http://dx.doi.org/10.1534/genetics.114.164350 (cf. p. 7, 8, 21).

Rakyan, Vardhman K. et al. (juil. 2011). "Epigenome-wide association studies for common human diseases". In : *Nature Reviews Genetics* 12.8, p. 529-541. ISSN : 1471-0064. DOI : 10.1038/nrg3000. URL : http://dx.doi.org/10.1038/nrg3000 (cf. p. 50).

Rañola, John Michael, John Novembre et Kenneth Lange (juil. 2014). "Fast spatial ancestry via flexible allele frequency surfaces". In : *Bioinformatics* 30.20, p. 2915-2922. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btu418. URL : http://dx.doi.org/10.1093/bioinformatics/btu418 (cf. p. 22).

Rellstab, Christian et al. (2015). "A Practical Guide To Environmental Association Analysis in Landscape Genomics". In : *Molecular Ecology* 24.17, p. 4348-4370. DOI : 10.1111/mec.13322. URL : https://doi.org/10.1111/mec.13322 (cf. p. 50).

Rubin, Donald B et al. (1981). "The bayesian bootstrap". In : *The annals of statistics* 9.1, p. 130-134 (cf. p. 52).

Simpson, Edward H (1951). "The interpretation of interaction in contingency tables". In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 238-241 (cf. p. 1).

Slonim, Donna K. (déc. 2002). "From patterns to pathways: gene expression data analysis comes of age". In : *Nature Genetics* 32.Supp, p. 502-508. ISSN : 1061-4036. DOI : 10.1038/ng1033. URL : http://dx.doi.org/10.1038/ng1033 (cf. p. 50).

Song, Minsun, Wei Hao et John D Storey (mar. 2015). "Testing for genetic associations in arbitrarily structured populations". In : *Nature Genetics* 47.5, p. 550-554. ISSN : 1546-1718. DOI : 10.1038/ng.3244. URL : http://dx.doi.org/10.1038/ng.3244 (cf. p. 67).

Stephens, Matthew (2016). "False Discovery Rates: a New Deal". In : *Biostatistics* nil.nil, kxw041. DOI : 10.1093/biostatistics/kxw041. URL : https://doi.org/10.1093/biostatistics/kxw041 (cf. p. 13).

Storey, J. D. (2011). "False Discovery Rate". In : *International Encyclopedia of Statistical Science*, p. 504-508. DOI : 10.1007/978-3-642-04898-2_248. URL : http://dx.doi.org/10.1007/978-3-642-04898-2_248 (cf. p. 73, 74).

Sun, Yunting, Nancy R. Zhang et Art B. Owen (déc. 2012). "Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene

expression data". In : *The Annals of Applied Statistics* 6.4, p. 1664-1688. ISSN : 1932-6157. DOI : 10.1214/12-aoas561. URL : http://dx.doi.org/10.1214/12-aoas561 (cf. p. 68).

TANG, Hua et al. (2005). "Estimation of individual admixture: Analytical and study design considerations". In : *Genetic Epidemiology* 28.4, p. 289-301. ISSN : 1098-2272. DOI : 10.1002/gepi.20064. URL : http://dx.doi.org/10.1002/gepi.20064 (cf. p. 7, 8, 21).

TIBSHIRANI, Robert (1996). "Regression shrinkage and selection via the lasso". In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 267-288 (cf. p. 58, 59).

TISHKOFF, Sarah A et al. (2009). "The genetic structure and history of Africans and African Americans". In : *science* 324.5930, p. 1035-1044 (cf. p. 6).

TSENG, P. (juin 2001). "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization". In : *Journal of Optimization Theory and Applications* 109.3, p. 475-494. ISSN : 1573-2878. DOI : 10.1023/a:1017501703105. URL : http://dx.doi.org/10.1023/a:1017501703105 (cf. p. 17, 60).

WANG, Jingshu et al. (2017). "Confounder adjustment in multiple hypothesis testing". In : *The Annals of Statistics* 45.5, p. 1863-1894 (cf. p. 13, 53, 70).

WEIR, B S (1996). *Genetic data analysis II: methods for discrete population genetic data*. Vol.2. Sinauer Associates, 445p. ISBN : 0878939024 (cf. p. 31).

WOLD, Svante (nov. 1978). "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models". In : *Technometrics* 20.4, p. 397-405. ISSN : 1537-2723. DOI : 10.1080/00401706.1978.10489693. URL : http://dx.doi.org/10.1080/00401706.1978.10489693 (cf. p. 30).

WOLLSTEIN, Andreas et Oscar LAO (2015). "Detecting individual ancestry in the human genome." In : *Investigative genetics* 6, p. 7. ISSN : 2041-2223. DOI : 10.1186/s13323-015-0019-x. URL : http://www.ncbi.nlm.nih.gov/pubmed/25937887%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4416275 (cf. p. 44).

WOLPERT, D.H. et W.G. MACREADY (avr. 1997). "No free lunch theorems for optimization". In : *IEEE Transactions on Evolutionary Computation* 1.1, p. 67-82. ISSN : 1089-778X. DOI : 10.1109/4235.585893. URL : http://dx.doi.org/10.1109/4235.585893 (cf. p. 88).

WRAY, Naomi R. et al. (juin 2013). "Pitfalls of predicting complex traits from SNPs". In : *Nature Reviews Genetics* 14.7, p. 507-515. ISSN : 1471-0064. DOI : 10.1038/nrg3457. URL : http://dx.doi.org/10.1038/nrg3457 (cf. p. 20).

YANG, Wen-Yun et al. (mai 2012). "A model-based approach for analysis of spatial structure in genetic data". In : *Nature Genetics* 44.6, p. 725-731. ISSN : 1546-1718. DOI : 10.1038/ng.2285. URL : http://dx.doi.org/10.1038/ng.2285 (cf. p. 22).

ZHOU, Xiang et Matthew STEPHENS (fév. 2014). "Efficient multivariate linear mixed model algorithms for genome-wide association studies". In : *Nature Methods* 11.4, p. 407-409. ISSN : 1548-7105. DOI : 10.1038/nmeth.2848. URL : http://dx.doi.org/10.1038/nmeth.2848 (cf. p. 12, 52).

ZHOU, Yan et al. (2016). "Sparse Multivariate Factor Analysis Regression Models and Its Applications To Integrative Genomics Analysis". In : *Genetic Epidemiology*

41.1, p. 70-80. DOI : 10.1002/gepi.22018. URL : https://doi.org/10.1002/gepi.22018 (cf. p. 53).

ZOU, James et al. (jan. 2014). "Epigenome-wide association studies without the need for cell-type composition". In : *Nature Methods* 11.3, p. 309-311. ISSN : 1548-7105. DOI : 10.1038/nmeth.2815. URL : http://dx.doi.org/10.1038/nmeth.2815 (cf. p. 74, 79, 82, 83, 89).

Travaux réalisés

Articles de revue

— **Kévin Caye**, Timo Deist, Helena Martins, Olivier Michel, Olivier François (2016) TESS3 : fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources* 16 (2), 540-548. DOI : 10.1111/1755-0998.12471.

— **Kévin Caye**, Flora Jay, Olivier Michel, Olivier Francois (2017). Fast Inference of Individual Admixture Coefficients Using Geographic Data. Accepted to *The Annals of Applied Statistics*. DOI : 10.1101/080291.

— Helena Martins, **Kévin Caye**, Keurcien Luu, Michael GB Blum, Olivier François (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology* 25 (20), 5029-5042. DOI : 10.1101/054585.

— Olivier François, Helena Martins, **Kévin Caye**, Sean D Schoville (2016). Controlling false discoveries in genome scans for selection. *Molecular ecology* 26 (2), 454-469. DOI : 10.1111/mec.13513.

Conférences

— **Kévin Caye**, Olivier Michel, Olivier François (2016). Algorithmes Pour l'Estimation des Coefficients de Métissage dans des Populations Continues Spatialement. 48èmes Journées de Statistique de la SFdS, 30 mai-3 juin 2016.

— **Kévin Caye**, Olivier Michel, Olivier François (2016). `tess3r` : étude du jeu de données Arabidopsis thaliana RegMap. Cinquièmes Rencontres R, 22-24 juin 2016.

— **Kévin Caye**, Olivier Michel, Olivier François (2016). `tess3r` : un package R pour l'estimation de la structure génétique des populations spatialisées. JOBIM,

28-30 juin 2016.

Logiciels

— **Kévin Caye**, Olivier François, Flora Jay. `tess3r` : An R package for estimating and visualizing spatial population structure based on geographically constrained non-negative matrix factorization and population genetics.

— **Kévin Caye**, Olivier François. `lfmm` : An R package for correcting association studies for confounding factors.

— Michael Blum, **Kévin Caye**, Thomas Dias Alves, Keurcien Luu. SSMPG2015 : Un site web pour l'organisation du challenge de l'école de printemps *Software and Statistical Methods for Population Genetics* 2015. https://ssmpg-challenge.imag.fr

— Michael Blum, **Kévin Caye**, Keurcien Luu, Florian Privé. SSMPG2017 : Une application web pour l'organisation du challenge de l'école de printemps *Software and Statistical Methods for Population Genetics* 2017. https://github.com/bcm-uga/SSMPG2017

Scripts

Les scripts utilisés pour réaliser les expériences présentées dans cette thèse sont disponibles à l'adresse : https://github.com/cayek/MaThese.