Modèles à facteurs latents pour les études d'association écologique en génétique des populations

Eric Frichot

▶ **To cite this version:**

Eric Frichot. Modèles à facteurs latents pour les études d'association écologique en génétique des populations. Médecine humaine et pathologie. Université de Grenoble, 2014. Français. NNT : 2014GRENS018 . tel-01557506

HAL Id: tel-01557506

https://theses.hal.science/tel-01557506v1

Submitted on 6 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

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

Arrêté ministériel : 7 août 2006

Présentée par

Éric FRICHOT

Thèse dirigée par **Olivier FRANÇOIS**
et codirigée par **Guillaume BOUCHARD**

préparée au sein **du laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG)**
et de **l'école doctorale "Ingénierie de la Santé, de la Cognition et Environnement" (EDISCE)**

Modèles à facteurs latents pour les études d'association écologique en génétique des populations

Thèse soutenue publiquement le **26 septembre 2014**,
devant le jury composé de :

Olivier Michel
Professeur, Grenoble INP, Président
Catherine Matias
DR CNRS, LPMA Paris, Rapporteur
Yves Vigouroux
DR IRD, Montpellier, Rapporteur
Olivier François
Professeur, Grenoble INP, Directeur de thèse
Guillaume Bouchard
CR Xerox, XRCE, Co-Directeur de thèse

Remerciements

Je souhaite tout d'abord remercier mon directeur de thèse, Olivier François, pour sa sympathie, son encadrement, son enseignement et sa patience. Si cette thèse s'est bien déroulée, c'est grâce à lui. Il a su se montrer toujours disponible pour discuter, répondre à mes questions, m'orienter vers des solutions et m'apprendre les nombreuses ficelles du métier d'enseignant-chercheur. Je remercie aussi mon co-directeur de thèse, Guillaume Bouchard, pour ses discussions et son ouverture d'esprit envers la génétique des populations. J'en profite pour remercier la région Rhône-Alpes pour l'allocation doctorale de recherche qui a été allouée pour ma thèse.

Je tiens aussi à remercier Catherine Matias, Yves Vigouroux et Olivier Michel, mon jury de thèse, pour s'être intéressés à ma thèse et pour avoir pris le temps de la réviser.

Je remercie aussi toute l'équipe BCM et plus généralement le laboratoire TIMC. Ca a été vraiment sympa de passer ces 3 ans avec eux. J'ai apprécié de toujours trouver du monde pour discuter de tous les sujets, ainsi que du monde pour toutes les activités de BCM comme le repas à la kfet, les pauses et les activités sportives avec le basket et l'escalade. En particulier, je tiens à remercier Nicolas Duforet-Frebourg, mon cobureau, pour sa bonne humeur. J'ai vraiment apprécié les discussions que nous avons pu avoir. J'aimerai aussi remercier toutes les personnes avec qui j'ai pu collaborer durant ma thèse. Ils ont tous contribué à l'élaboration de ce travail.

Enfin, je remercie tous mes amis et ma famille pour leur soutien et leur présence à ma soutenance de thèse. En particulier, je remercie mes colocs pour ces 3 années de vie en communauté. Ca a été une expérience inoubliable que s'est très bien équilibrée avec le travail de thèse. Pour finir, je tiens à remercier Marine. Elle a su m'apporter son soutien et un équilibre indispensable qui m'ont permis de donner le meilleur de moi-même dans cette thèse.

Titre : Modèles a facteurs latents pour les études d'association écologique en génétique des populations.

Résumé : Nous introduisons un ensemble de modèles à facteurs latents dédié à la génomique du paysage et aux tests d'associations écologiques. Cela comprend des méthodes statistiques pour corriger des effets d'autocorrélation spatiale sur les cartes de composantes principales en génétique des populations (spFA), des méthodes pour estimer rapidement et efficacement les coefficients de métissage individuel à partir de matrices de génotypes de grande taille et évaluer le nombre de populations ancestrales (sNMF) et des méthodes pour identifier les polymorphismes génétiques qui montrent de fortes corrélations avec des gradients environnementaux ou avec des variables utilisées comme des indicateurs pour des pressions écologiques (LFMM). Nous avons aussi développé un ensemble de logiciels libres associés à ces méthodes, basés sur des programmes optimisés en C qui peuvent passer à l'échelle avec la dimension de très grand jeu de données, afin d'effectuer des analyses de structures de population et des cribles génomiques pour l'adaptation locale.

Mots-clés : modèles à facteurs latents, adaptation locale, structure génétique des populations, séquencage haut-debit, statistiques bayésiennes, apprentissage.

Title : Latent factor models for ecological association studies in population genetics.

Summary : We introduce a set of latent factor models dedicated to landscape genomics and ecological association tests. It includes statistical methods for correcting principal component maps for effects of spatial autocorrelation (spFA) ; methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (sNMF) ; and methods for identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures (LFMM). We also developed a set of open source softwares associated with the methods, based on optimized C programs that can scale with the dimension of very large data sets, to run analyses of population structure and genome scans for local adaptation.

Keywords : latent factor models, local adaptation, population genetic structure, next generation sequencing, bayesian statistics, machine learning.

Table des matières

Résumé de la thèse

On parle d'adaptation locale d'une espèce à son environnement pour décrire l'ajustement fonctionel des organismes de cette espèce à son habitat. L'objectif de cette thèse est de développer des modèles statistiques pour détecter des signaux d'adaptation locale au sein des génomes d'une espèce par criblage génomique. De nombreuses méthodes ont préalablement été développées avec cet objectif. Toutefois, ces méthodes ont des difficultés pour prendre en compte les facteurs de confusion propres à la génétique des populations. Les facteurs de confusion peuvent être, par exemple, créés par l'histoire démographique de l'espèce étudiée, la répartition spatiale des échantillons ou bien l'existence de facteurs environnementaux dont on n'aurait pas tenu compte. En effet, il est difficile de distinguer la variation génétique due à l'adaptation locale de la variation génétique due à la dérive génétique et à l'histoire démographique.

Dans la première partie de la thèse, nous nous sommes intéressés à la modélisation de la structure génétique des populations grâce à des modèles à facteurs latents. Un ensemble d'individus est dit génétiquement structuré s'il existe des populations homogènes au sein de cet ensemble. De nombreux phénomènes, liés à l'histoire ou aux pressions environnementales, peuvent entraîner l'existence de structure génétique dans un groupe d'individus. Il existe deux types principaux d'approches pour étudier la structure génétique des populations : les approches de type "analyse en composantes principales" où l'on cherche à résumer la variation génétique entre individus en un nombre réduit de composantes et les approches de type "structure" où l'on cherche à évaluer les proportions de métissage de chaque individu.

Tout d'abord, nous avons développé le modèle spFA, pour l'analyse factorielle spatiale, une extension de l'analyse en composantes principales prenant explicitement en compte la répartition géographique des individus afin de modéliser l'isolement par la distance. Nous avons montré que le modèle spFA est capable de produire des facteurs principaux corrigeant les formes sinusoïdales que l'on observe dans l'analyse en composantes principales en cas d'isolement par la distance. Cela facilite l'interprétation des facteurs principaux et par conséquent, facilite l'étude de la structure génétique des populations. Ce travail a fait l'objet d'une publication intitulée "Correcting principal component maps for effects of spatial autocorrelation in population genetic data" (Frontiers in Genetics. 2012 ; 3 :254).

Ensuite, nous nous sommes intéressés à l'estimation de coefficients de métissage individuel pour des jeux de données à haute résolution génomique comportant des millions de SNPs. Afin de construire un modèle flexible, nous avons proposé une nouvelle méthode d'estimation des coefficients de métissage fondée sur la décomposition en facteurs de la

matrice de génotypes. Cette approche s'appuie sur l'unification des approches de type "analyse en composantes principales"et "structure" grâce à une factorisation de matrice non négative parcimonieuse (sNMF). Le logiciel proposé, `sNMF`, est 10 à 30 fois plus rapide que le logiciel servant d'état-de-l'art sur les jeux de données que nous avons analysés. De plus, nous avons proposé un critère prédictif pour déterminer le nombre de populations ancestrales. Notre modélisation permet d'analyser des données issues d'espèces consanguines sans restriction particulière. Nous avons illustré notre approche par l'étude de plusieurs jeux de données humaines et d'un jeu de données de la plante *Arabidopsis thaliana* en Europe. Ce travail a fait l'objet d'une publication intitulée "Fast and efficient estimation of individual ancestry coefficients" (Genetics. 2014 ; 196 :973–983).

Dans la seconde partie de la thèse, nous nous sommes intéressés au développement de modèles statistiques pour détecter l'adaptation locale au sein des génomes d'une espèce par criblage génomique grâce à des modèles à facteurs latents. Nous avons proposé de nouveaux algorithmes fondés sur la génétique des populations, la modélisation en écologie et des techniques d'apprentissage statistique afin de cribler des génomes à la recherche de signatures d'adaptation locale. Implantés dans le programme `LFMM`, les modèles mixtes à facteurs latents utilisent une approche dans laquelle la structure de population est modélisée par des variables non-observées. Des algorithmes rapides et efficaces détectent des corrélations entre des variations génétiques et environnementales tout en estimant simultanément les différents niveaux de structure de population. Comparer ces algorithmes avec des méthodes similaires nous a montré que le logiciel `LFMM` peut estimer efficacement les effets aléatoires liés à l'histoire des populations et aux patrons d'isolement par la distance lorqu'il estime les corrélations entre gènes et environnement. De plus, le logiciel `LFMM` diminue le nombre de fausses associations dans des études de criblage génomique. Nous avons appliqué le logiciel `LFMM` à des jeux de données humaines et à des plantes, en mettant en avant plusieurs gènes ayant des fonctions associées au développement et montrant de fortes corrélations avec des gradients climatiques. Ce travail a fait l'objet d'une publication intitulée "Testing for associations between loci and environmental gradients using latent factor mixed models" (Molecular Biology and Evolution. 2013 ; 30 :1687–1699).

Mots-clés
adaptation locale, modèles à facteurs latents, structure génétique des populations.

Chapitre 1

État de l'art

L'objectif de ce chapitre introductif est de décrire le contexte de la thèse et l'état de l'art des méthodes élaborées dans le manuscrit. Pour cela, nous rappelons les concepts nécessaires afin d'introduire la notion d'adaptation locale à un environnement. Puis, nous faisons l'état de l'art des méthodes existantes pour détecter l'adaptation locale d'une espèce à un environnement à partir de données génomiques obtenues pour des individus échantillonnés dans leur habitat naturel. Nous clôturons ce chapitre en détaillant les objectifs de cette thèse.

1.1 Contexte de la génomique des populations

Nous rappelons ici les motivations de la génomique des populations en liaison avec l'écologie et plus particulièrement de la génomique du paysage. Puis, nous détaillons le type de données auxquelles nous nous sommes intéressés et les données que nous avons utilisées. Enfin, nous présentons les concepts de sélection naturelle et d'adaptation locale utilisés dans ce travail.

1.1.1 Motivations

La complexité des processus responsables de la variation phénotypique recouvre de vastes échelles, allant de l'ADN, soumis à des processus moléculaires de mutation ou de recombinaison, jusqu'aux populations évoluant selon des processus démographiques et migratoires. La prise en compte de toutes ces échelles a donné naissance au domaine encore

émergent de la génomique des populations, cherchant en particulier, dans chaque population, à caractériser au niveau moléculaire les gènes soumis à la sélection naturelle.

La génomique des populations s'intéresse à l'étude des variations du génome d'un ensemble d'individus issus de populations d'une même espèce. Ces variations sont le reflet de l'évolution de l'espèce étudiée (Darwin, 1859; Williams, 1966). Elles permettent de mieux comprendre l'histoire d'une espèce, l'origine de la spéciation, les migrations, la variation des tailles de populations, et les pressions évolutives que peut subir cette espèce (Luikart et al., 2003).

La génétique du paysage est née de la fusion des concepts de l'écologie du paysage et de la génétique des populations (Manel et al., 2003). La génétique du paysage s'attache à décrire comment la structure génétique des populations peut être influencée par la structure du "paysage" et comment des modifications environnementales sont susceptibles d'impacter la diversité génétique des populations étudiées. Les étapes importantes de la génétique du paysage sont, d'une part, la détection des discontinuités génétiques comme par exemple des barrières aux flux de gènes au sein des populations et d'autre part, la corrélation de telles discontinuités avec les variables environnementales et écologiques (Manel et al., 2003; Duforet-Frebourg and Blum, 2014).

Suite au développement des technologies de séquençage et à l'apparition du domaine de la génomique des populations, la discipline de la génomique du paysage a émergé plus récemment (Joost et al., 2007). La génomique du paysage donne un cadre pour l'étude des variations génétiques neutres et adaptatives à l'échelle des populations dans un contexte spatial (Barrett and Hoekstra, 2011; Schoville et al., 2012). Ce cadre a pour but de regrouper des données environnementales, des échantillons à très grande résolution de séquençage des génomes, et des méthodes statistiques afin de mieux comprendre l'écologie des espèces et l'adaptation écologique (Manel et al., 2010; Schoville et al., 2012). La génomique du paysage possède de nombreuses applications, allant de la biologie de la conservation à l'épidémiologie moléculaire (Segelbacher et al., 2010).

L'objectif de la thèse est de construire de nouvelles méthodes statistiques pour la génomique du paysage. D'une part, le but est de mieux estimer la structure génétique des individus au sein d'une espèce. D'autre part, le but est de détecter les corrélation entre les variations génétiques au sein d'une espèce et la structure du paysage, afin de rechercher des bases moléculaires à l'adaptation des espèces. On appelle cela *une étude d'association écologique*.

1.1.2 Données de polymorphismes génétiques

Les variations au sein des génomes d'une espèce sont appelées *des polymorphismes génétiques*. Les polymorphismes génétiques s'accumulent au cours du temps à travers des processus de mutation, et sont de différents types. Cela peut être l'insertion, la délétion ou l'échange d'un ou de plusieurs nucléotides dans le génome. Dans le cadre de cette thèse, nous nous intéressons aux polymorphismes d'un seul nucléotide appelés *SNPs*, pour "Single Nucleotide Polymorphisms". Il existe alors plusieurs variants du polymorphisme, appelés *allèles*, pour un locus donné, un *locus* étant un emplacement physique sur le génome. L'allèle existant avant la mutation est appelé *allèle de référence*, *allèle ancestral* ou *allèle parent*. L'allèle résultant de la mutation est appelé *allèle muté*, ou *dérivé*. Le processus technologique permettant de déterminer les allèles spécifiques à chaque individu d'une espèce est appelé *le génotypage*.

La croissance rapide de la génétique des populations est en partie due au développement d'outils puissants permettant le génotypage de nombreux individus à un grand nombre de locus, et de projets internationaux. Par exemple, le projet HGDP pour "Human Genome Diversity Project" a vu le jour à Stanford en collaboration avec le Centre d'Étude du Polymorphisme Humain à Paris (Rosenberg et al., 2002; Cavalli-Sforza, 2005; Li et al., 2008). Grâce à la puce Illumina 650Y, 1043 humains répartis dans 51 populations à travers le monde, ont été génotypés à plus de 650,000 locus. D'autre projets d'étude de l'espèce humaine, tel que HapMap, ont vu le jour (Gibbs et al., 2003). L'objectif du projet HapMap est de développer une carte génétique qui décrit les patrons communs de variations du génome humain. Des projets sont aussi en cours de développement pour le génotypage de l'espèce de plante *Arabidopsis thaliana* (Arabidopsis Genome Initiative., 2000; Weigel and Mott, 2009). Un but est de mieux comprendre les pressions de sélection qui s'exercent sur cette plante que l'on peut trouver partout sur la planète.

Par la suite, la possibilité de séquencer des génomes entiers dans un délai très court et à un coût raisonnable a offert des perspectives sans précédent pour la biologie évolutive et la médecine personnalisée (1000 Genomes Project Consortium., 2010). Par exemple, le projet du 1000 Genomes Project Consortium est de séquencer 2577 génomes complets d'individus répartis à travers le monde. Les chercheurs espèrent ainsi mieux comprendre les corrélations entre le profil génétique d'un organisme vivant, le génotype, et la caractéristique d'évolution, le phénotype.

L'analyse des jeux de données qui résultent de telles études de génotypage est la source de problèmes nouveaux pour les modélisateurs en génomique des populations. Ces nouvelles questions sont posées à la fois par la dimension des données et par l'accroissement de la complexité des modèles considérés. La nature des données acquises par les nouvelles

technologies de séquençage nécessite en toute évidence de concevoir de nouvelles méthodes d'analyse statistique afin de pouvoir synthétiser l'information utile aux biologistes.

1.1.3 Exemples d'étude d'association écologique pour détecter l'adaptation locale

Dans son livre "L'origine des espèces", Darwin pose les fondements de la théorie de l'évolution et de la sélection naturelle. Il y explique que la variabilité individuelle est le fondement de l'évolution par le moyen de la sélection naturelle. A cause des pressions environnementales sélectives, les lignées des individus les plus adaptés à leur environnement ont plus de chance de perdurer que les lignées moins adaptées (Darwin, 1859; Williams, 1966).

Dans de nombreux exemples, l'environnement varie au sein de l'espace géographique de l'espèce considérée. Un individu, localement moins adapté à un environnement donné peut être plus adapté à un environnement différent. On parle d'adaptation locale à un environnement. L'adaptation locale à travers la sélection naturelle joue donc un rôle essentiel dans la construction des variations au sein de populations.

De nombreux exemples d'adaptation locale ont pu être observés dans la nature. Par exemple, Simonson et al. (2010) ont étudié l'adaptation des tibétains à l'altitude. Les tibétains vivent à très haute altitude depuis des milliers d'années. Ils ont un ensemble de traits physiques qui leur permet de mieux tolérer l'hypoxie. Ces traits sont clairement le résultat d'une adaptation à leur environnement. À l'aide d'une étude de criblage génomique, Simonson et al. (2010) ont déterminé un ensemble de régions génomiques contenant des gènes qui semblent impliqués dans l'adaptation à l'altitude. En particulier, les gènes EGLN1 et PPARA ont été significativement associés avec une diminution de l'hémoglobine dans le sang, phénomène caractéristique des populations vivant en haute altitude.

D'autres exemples naturels d'adaptation locale ont pu être observés. Chez l'espèce humaine, Hancock et al. (2011) ont étudié l'adaptation aux pressions climatiques, tandis que Fumagalli et al. (2011) ont mis en avant l'importance de l'adaptation à plusieurs pathogènes comme facteur de sélection. Parmi les études d'association écologique pour des espèces modèles animales et végétales, Hancock et al. (2011) ont mis en évidence des signatures de l'adaptation locale dans le génome de la plante *Arabidopsis thaliana*, tandis que Hohenlohe et al. (2010) ont détecté l'adaptation parallèle chez l'espèce modèle de petit poisson épinoche, *Gasterosteus aculeatus*. Des études ont aussi été réalisées sur

des espèces non-modèles. Eckert et al. (2010) ont, par exemple, pointé l'adaptation locale chez l'espèce de pins *Pinus taeda*, Pinaceae.

Ces études récentes montrent que la détection d'adaptation locale est une problématique importante actuellement. Il est donc important de développer des méthodes fiables pour détecter l'adaptation locale au moyen de criblage génomique afin de minimiser les faux positifs dans les tests d'association écologique.

1.2 Détection de l'adaptation locale par criblage génomique

L'adaptation locale à un environnement est un phénomène essentiel pour mieux comprendre l'évolution d'une espèce à son paysage. De plus, les récents développements en séquençage génétique apportent de nouvelles informations à traiter pour comprendre ces processus. Nous proposons dans cette partie de décrire les différentes approches pour réaliser une étude d'association écologique afin détecter de l'adaptation locale au sein des génomes. Pour cela, nous présentons le test de Mantel, puis les modèles linéaires. Parmi les modèles linéaires, nous nous intéressons à la régression linéaire, aux modèles linéaires généralisés, à la régression linéaire avec bruit corrélé, au modèle du logiciel `BAYENV` et aux modèles mixtes. Nous discutons, pour chaque méthode les hypothèses statistiques et génétiques sous-jacentes aux modèles.

1.2.1 Données et notations

Nous définissons des notations communes à l'ensemble des méthodes. Nous considérons une matrice $(G_{i\ell})$, où chaque entrée est la fréquence d'allèle ou génotype pour l'individu i au locus ℓ. Le nombre d'individus est noté n et le nombre de locus est noté L. Pour plus de simplicité, on suppose que les locus sont bialléliques, des SNPs. Dans ce cas, $G_{i\ell}$ est le nombre d'allèles dérivés au locus ℓ pour l'individu i. $G_{i\ell}$ peut valoir 0, 1 ou 2 dans le cas d'individus diploides. On dit que l'individu i au locus ℓ est homozygote pour l'allèle ancestral lorsque $G_{i\ell} = 0$, hétérozygote lorsque $G_{i\ell} = 1$ et homozygote pour l'allèle dérivé lorsque $G_{i\ell} = 2$. Nous notons $G^{(c)}$, la matrice de génotypes centrée à chaque locus. La matrice $G^{(c)}$ s'écrit

$$G^{(c)}_{i\ell} = G_{i\ell} - \mu_\ell,$$

où μ_ℓ est la moyenne des génotypes pour le locus ℓ. De même, nous notons $G^{(n)}$ la matrice des génotypes normée pour chaque locus. La matrice $G^{(n)}$ s'écrit

$$G^{(n)}_{i\ell} = \frac{G_{i\ell} - \mu_\ell}{\sqrt{f_\ell(1-f_\ell)}},$$

où $f_\ell = \frac{\mu_\ell}{2}$ est la moyenne de la fréquence d'allèle 1 au locus ℓ. Cette normalisation se justifie par le fait que pour un marqueur biallélique, $f_\ell(1-f_\ell)$ représente la variance du nombre d'allèles 1 au locus ℓ dans le cas d'un équilibre d'Hardy-Weinberg (Patterson et al., 2006). À ces données génotypiques, nous associons un ensemble de d variables géographiques ou environnementales pour chaque individu i, représentées par le vecteur X_i.

Dans cette partie, consacrée à la détection de l'adaptation locale, nous utilisons la matrice de génotypes centrée, $G^{(c)}$, pour simplifier les notations. Dans certains cas, il a été montré que l'intégration de la moyenne pour chaque SNP apporte plus de flexibilité au modèle statistique (Engelhardt and Stephens, 2010). On pourra facilement généraliser les modèles présentés en ajoutant l'estimation de la moyenne de chaque SNP. Les modèles peuvent sans difficulté considérer des microsattelites, des polymorphismes de longueur des fragments amplifiés (AFLPs) ou des données de comptage d'allèles.

En ce qui concerne les notations mathématiques, nous utilisons une minuscule pour faire référence à un scalaire, tandis que nous utilisons une lettre majuscule pour faire référence à un vecteur ou à une matrice. On note $N(\mu, \Sigma)$ la loi normale multivariée de moyenne μ et de matrice de covariance Σ, $\Gamma(a, b)$, la loi gamma de forme a et d'intensité b et χ^2_d la loi du χ^2 à d degrés de liberté. On ajoute un accent circonflexe à la variable pour désigner son estimateur. Par exemple, l'estimateur de μ_ℓ, l'espérance des génotypes pour le locus ℓ, est noté $\hat{\mu}_\ell$. On note $E(x)$ l'espérance de x, $var(x)$ la variance de x et $cov(Y)$ la matrice de covariance associée au vecteur Y.

1.2.2 Indicateurs environnementaux

Pour détecter les corrélations entre des SNPs et des pressions environnementales, il faut représenter ces pressions sous forme de variables quantitatives. On considère donc une variable ou un ensemble de variables environnementales représentatives de la pression adaptative que l'on cherche à étudier. Comme ces variables ne représentent pas directement l'action de la pression mais plutôt les facteurs observés de cette pression, on parle d'*"indicateurs" environnementaux*. Pour donner un exemple, on peut considérer des variables de température ou de précipitation pour représenter des pressions climatiques

(Eckert et al., 2010). Dans le cadre de l'étude de l'adaptation d'un pathogène à son hôte, on peut considérer la diversité génétique du pathogène comme facteur d'adaptation. En effet, on peut faire l'hypothèse qu'un pathogène ayant une plus grande diversité génétique aura une meilleure capacité à s'adapter aux différentes réponses de l'hôte à la présence du pathogène (Hancock et al., 2008).

1.2.3 Test de Mantel

Une des méthodes les plus connues pour déterminer la relation entre des génotypes et un indicateur environnemental est le test de Mantel (Mantel, 1967). L'objectif de ce test est de calculer la corrélation entre deux matrices afin d'évaluer si cette corrélation est significative en la comparant à la loi des valeurs obtenues à la suite de permutations au sein de ces deux matrices (Mantel, 1967). Le test de Mantel a rapidement été utilisé en génétique des populations pour tester l'absence de corrélation entre une matrice contenant les distances génétiques entre individus et une matrice contenant les distances géographiques ou environnementales entre ces mêmes individus. Développé plus récemment, le test de Mantel partiel permet d'évaluer l'effet d'une variable sur une autre, tout en contrôlant l'effet d'une troisième (Smouse et al., 1986). Depuis, il a souvent été appliqué en génétique des populations, notamment à des données de populations humaines pour tester la corrélation entre distance génétiques et diversité des pathogènes (Fumagalli et al., 2011). Fumagalli et al. (2011) corrèlent la répartition spatiale des fréquences d'allèles d'un grand nombre de SNPs chez 55 populations humaines avec des facteurs environnementaux, tels que le climat et la diversité génétique des pathogènes. Les auteurs utilisent un test de Mantel partiel corrigé pour la démographie. Ils déterminent 103 gènes fortement associés avec l'environnement pathogénique. En particulier, ils associent à la diversité pathogénique 34 gènes qui ont été associés, dans des études précédentes, à la vulnérabilité à des maladies auto-immunes (Table 1.1).

Plusieurs difficultés sont néanmoins présentes lors de l'utilisation des tests de Mantel. Tout d'abord, le choix de la distance génétique ou environnementale peut influencer le résultat du test. Belle and Barbujani (2007) ont montré que les conclusions de leur étude sur le rôle de la géographie et de la langue dans la distribution de la diversité génétique différaient selon qu'ils appliquaient leurs tests à des distances génétiques mesurées par l'indice de différenciation F_{ST} (mesure la plus couramment utilisée) ou par l'indice R_{ST} (Slatkin, 1995). De plus, les tests de Mantel nécessitent de choisir quelle technique de permutation permettra d'obtenir la loi des corrélations neutres correspondant à l'hypothèse nulle du test (Legendre, 2000). Enfin, Frichot et al. (2013) ont montré, sur des simulations mimant la structure de population, que le test de Mantel n'était pas correctement

Disease	Associated genes
Celiac disease	*TNFRSF9, CCR4, ETS1, CD28, SH2B3, REL, CIITA, LPP, SLC9A4*
Ulcerative colitis	*IL10, IFNG, CIITA, FCGR2A, IL19, REL*
Type 1 diabetes	*DLK1, SH2B3, CTSH, IL10, C16orf75*
Multiple sclerosis (susceptibility, severity or age of onset)	*CBLB, RPL5, KCNB2, CENPC1, FUT8*
Systemic lupus erythematosus	*ETS1, SOCS6*
Rheumatoid arthritis	*REL, SH2B3*
Vitiligo	*LPP, RERE*
Crohn's disease (or combined with sarcoidosis)	*FUT2*
Behcet's disease	*IL10*
Ankylosing spondylitis	*ANTXR2*

doi:10.1371/journal.pgen.1002355.t003

TABLE 1.1: Liste de gènes associés à la diversité des pathogènes précédemment mis en avant par d'autres études pour être corrélés à la vulnérabilité face à des maladies auto-immunes (Fumagalli et al., 2011).

calibré. Ces difficultés limitent l'application en pratique du test de Mantel en génétique des populations.

1.2.4 Modèles de régression linéaire et généralisations

Une autre façon de chercher des signatures d'adaptation locale, en particulier lorsque les allèles avantageux ont un effet phénotypique faible, est d'identifier les SNPs qui montrent une forte corrélation avec des variables environnementales (Joost et al., 2007; Hancock et al., 2008; Coop et al., 2010; Poncet et al., 2010; Pritchard et al., 2010). Dans la nature, les traits quantitatifs qui montrent une variation géographique continue sont souvent associés à des variables écologiques spécifiques (Endler, 1977). Ces indicateurs écologiques reflètent les pressions sélectives agissant sur les phénotypes des individus. La variation des traits dans l'espace se reflète dans des clines géographiques ou dans des populations sympatriques qui exploitent différentes niches écologiques (Haldane, 1948; Berry and Kreitman, 1993; Prugnolle et al., 2005; Young et al., 2005). Par conséquence, l'adaptation locale à des environnements continus peut être détectée si il y a une association avec des variables environnementales à certains locus significativement supérieure à celle trouvée dans la variation génomique de fond.

1.2.4.1 Régression linéaire

La régression linéaire est une méthode simple pour déterminer une corrélation entre l'environnement et la variation génétique. Dans cette partie, nous présentons la régression linéaire d'un point de vue statistique puis nous présentons ces variantes, son application

en génétique des populations ainsi que ses limites. Nous profitons de ce modèle simple pour introduire des concepts communs à toutes les approches corrélatives, tels que les tests, la correction pour les tests multiples, le contrôle des fausses découvertes.

Le modèle sous-jacent à la régression linéaire s'écrit de la manière suivante :

$$G_{i\ell}^{(c)} = X_i B_\ell^T + \epsilon_{i\ell}, \quad i = 1 \dots n, \ell = 1 \dots L,$$

où $\epsilon_{i\ell}$ est un résidu de loi $N(0, \sigma_\ell^2)$ et B_ℓ est le vecteur contenant les coefficients de régression pour le locus ℓ, de dimension $d \geq 1$.

L'estimateur du vecteur des coefficients de régression au locus ℓ est donné par la formule suivante :

$$\hat{B}_\ell = (X^T X)^{-1} X^T G_\ell^{(c)}.$$

L'estimateur du vecteur des coefficients de régression au locus ℓ est sans biais ($E(\hat{B}_\ell) = B_\ell$) et de variance minimale ($cov(\hat{B}_\ell) = \hat{\sigma_\ell^2}(X^T X)^{-1}$).

1.2.4.2 Tests statistiques

La régression linéaire permet de chercher deux types d'effets, les effets forts et les effets significatifs. En effet, si on regarde l'estimation ponctuelle du coefficient de régression, $B_{\ell j}$ au locus ℓ pour l'indicateur environnemental j, on peut déterminer si l'effet associé à ce coefficient est fort en regardant si la valeur de $B_{\ell j}$ est grande positivement ou négativement. En génétique des populations, on s'intéresse aussi à la détection d'effets faibles mais significativement associés avec l'indicateur environnemental. Ainsi, on peut déterminer les corrélations les plus significatives en effectuant un test. Pour la régression linéaire, on peut tester l'hypothèse "$B_{\ell j} = 0$" à l'aide de la loi de Student à $n-d$ degrés de liberté, où d est le nombre de variables environnementales. En pratique, si on a suffisamment d'individus (≥ 40), la loi de Student s'approche très bien par une loi normale. Alternativement au t-test (test utilisant la loi de Student), on peut aussi effectuer un test de Wald, utilisant une loi du χ^2.

Pour effectuer un test, on peut calculer un z-score. Si l'on suppose qu'une variable aléatoire $B_{\ell j}$ est d'espérance $E(B_{\ell j})$ et d'écart type $\sigma(B_{\ell j})$, le calcul d'un z-score s'effectue de la manière suivante :

$$z = \frac{B_{\ell j} - E(B_{\ell j})}{\sigma(B_{\ell j})}.$$

On obtient finalement une p-valeur qui représente la probabilité de rejeter à tort l'hypothèse nulle ("$B_{\ell j} = 0$" dans le cas de la régression linéaire).

Alternativement au z-score, on peut effectuer un test de rapport de vraisemblances. Le but de ce test est de comparer l'apport du modèle alternatif par rapport au modèle nul. On considère donc le modèle nul, où "$B_{\ell j} = 0$", à d_1 degrés de liberté et le modèle où "$B_{\ell j} \neq 0$", à d_2 degrés de liberté. On calcule alors la statistique D,

$$D = -2 \log \left(\frac{\text{vraisemblance du modèle nulle}}{\text{vraisemblance du modèle alternatif}} \right)$$

et on utilise pour statistique de test, une loi du $\chi^2_{d_2 - d_1}$.

En statistique bayésienne, on s'intéresse au facteur de Bayes plutôt qu'au rapport de vraisemblances. On considère ainsi deux modèles M_1 et M_2. Le facteur de Bayes est le rapport des lois marginales sachant le modèle M_1 et sachant le modèle M_2. On considère que le modèle M_1 est plus probable que le modèle M_2 si le facteur de Bayes est supérieur à 1. Dans le cas d'une régression linéaire, on peut calculer le facteur de Bayes pour le modèle "$B_{\ell j} \neq 0$" et le modèle "$B_{\ell j} = 0$".

En effectuant les tests d'association en génomique des populations, on associe à chaque SNP une p-valeur. On considérera qu'un SNP est significativement associé à un indicateur environnemental si la p-valeur associée est plus petite qu'un seuil de significativité, α. Il faut alors déterminer le seuil α. Il dépend de nombreux paramètres, tels que le nombre de tests effectués ou la taille des échantillons. Il est important d'en discuter le choix (Johnson, 2013). En génétique des populations, la difficulté réside dans le fait que l'on réalise de nombreux tests puisque l'on effectue un test par SNP (pour plusieurs millions de SNPs) et par indicateur environnemental. Ainsi, il est nécessaire de choisir le seuil α afin de prendre en compte les tests multiples.

La correction la plus simple pour les tests multiples est la correction de Bonferroni. Elle consiste approximativement à diviser le seuil α par le nombre de tests que l'on effectue (Dunn, 1961). Cette correction fonctionne bien si les tests sont indépendants, c'est-à-dire si les SNPs sont indépendants les uns des autres. Hors, en génétique des populations, les SNPs sont corrélés par ce que l'on appelle *le déséquilibre de liaison* (Nyholt, 2004). Le déséquilibre de liaison est le fait que des SNPs physiquement proches peuvent être corrélés. Dans ce cas, la correction de Bonferroni a tendance à être trop conservative, c'est-à-dire que cette correction propose une seuil trop petit par rapport au seuil à appliquer pour corriger pour les tests multiples.

Alternativement à la p-valeur, on peut chercher à contrôler le taux de Fausses Découvertes (FDR). Pour cela, on peut utiliser la procédure de Benjamini-Hochberg (Benjamini and Hochberg, 1995) ou estimer une q-valeur (Storey, 2002). Le contrôle du taux de fausses découvertes consiste à maitriser la proportion de faux positifs parmi les positifs. Plus précisément, la q-valeur est le taux de fausses découvertes minimal pour lequel le test est significatif. La q-valeur peut être calculée à partir d'une loi empirique de p-valeurs (Storey, 2002). En génétique des populations, l'objectif du taux de fausses découvertes est de retourner une liste d'associations significatives la plus grande possible (puissance maximale) tout en contrôlant le nombre d'erreurs dans cette liste (FDR contrôlé).

1.2.4.3 Variantes de la régression linéaire

Joost et al. (2007) proposent d'utiliser un modèle de régression logistique afin de mieux prendre en compte les données binaires, de type absence ou présence d'un marqueur, dans le logiciel `SAM` (Spatial Analysis Methods). Ils proposent aussi d'ajouter les coordonnées géographiques à l'indicateur environnemental afin de considérer une correction pour la structure des populations.

Dans ce cadre, ils considèrent des données haploides. Grâce à une régression logistique, ils modélisent la fréquence de l'allèle 1 au locus ℓ pour l'individu i, $f_{i\ell}$, par une fonction de lien logit

$$\ln \frac{p(f_{i\ell}|X_i)}{1-p(f_{i\ell}|X_i)} = X_i B_\ell^T + \epsilon_{i\ell},$$

où $p(f_{i\ell}|X_i)$ est la probabilité de présence a posteriori de l'allèle muté au locus ℓ chez l'individu i et $\epsilon_{i\ell}$ est un résidu de loi $N(0, \sigma_\ell^2)$.

De manière plus globale, on peut considérer l'utilisation de modèles linéaires généralisés (GLM, McCullagh and Nelder (1989)). Le principe des GLM est d'inclure une fonction de lien entre les données et la variable explicative dans le modèle de régression.

Joost et al. (2007) appliquent le logiciel `SAM` à l'espèce de charançon *Hylobius abietis* L.(Curculionidas), l'un des insectes les plus économiquement nuisibles dans les forêts européennes de conifères. Leur échantillon est composé de 367 individus en Europe pour 83 génotypes de type AFLP. Les indicateurs environnementaux sont l'altitude et un ensemble de variables climatiques regroupant des informations sur la quantité de précipitation et la température. Cette étude permet de déterminer un ensemble de 11 marqueurs significativement associés avec leurs indicateurs environnementaux et dont les fonctions restent à déterminer.

En conclusion, le fait de prendre en compte la spécificité des données binaires propre à la génétique des populations afin de mieux modéliser la variation génétique a permis de détecter des nouvelles associations avec des facteurs environnementaux.

1.2.4.4 Facteurs de confusion

L'adaptation locale à des environnements continus peut être détectée s'il y a une association significative avec des variables environnementales à certains locus en comparaison de la variation génomique de fond. Une des difficultés principales est que la répartition géographique commune des variations environnementales et génétiques peut fausser les associations (Eckert et al., 2010). Par exemple, l'adaptation locale peut être entravée par des flux de gènes (Lenormand, 2002) et il peut être difficile de distinguer les effets dus à l'adaptation locale des effets dus à la dérive génétique et à l'histoire démographique. Les facteurs de confusion peuvent être, par exemple, créés par l'histoire démographique de l'espèce étudiée, la répartition spatiale des échantillons ou bien l'existence de facteurs environnementaux dont on n'aurait pas tenu compte (Novembre and Di Rienzo, 2009).

Sans correction pour les effets dus à la structure des populations ou à l'isolement par la distance, le modèle nul de la régression linéaire est incorrect pour prendre en compte l'histoire démographique de l'organisme étudié (Frichot et al., 2013). Meirmans (2012) a montré que des tests d'association entre des locus et des variables environnementales utilisant des modèles de régression linéaire classiques (GLM) produisent un fort taux de fausses associations ou fausses découvertes. Enfin, les pressions évolutives sont souvent faibles, polygéniques et réparties sur tout le génome. Cela a pour conséquence de noyer les véritables corrélations environnementales parmi les fausses associations.

Afin de réaliser une étude d'association écologique avec des méthodes corrélatives, il est donc nécessaire de corriger les facteurs de confusion propres à la génétique des populations.

1.2.5 Approches prenant en compte les facteurs de confusion

Différentes approches ont été proposées pour prendre en compte les facteurs de confusion et ainsi modéliser avec plus de précision les corrélations entre locus et indicateurs environnementaux. Une façon de procéder est d'estimer une structure génétique dite "neutre" pour l'échantillon, c'est-à-dire indépendante de toute pression de sélection, puis de corriger les méthodes de corrélation pour cette structure entre individus. Dans cette partie, nous présentons la régression linéaire à bruit corrélé, le modèle du logiciel `BAYENV` et les modèles mixtes.

1.2.5.1 Modèles de régression linéaire à bruit corrélé

La régression linéaire à bruit corrélé est une façon de modéliser la corrélation entre une variable environnementale et la variation génétique tout en prenant en compte la structure "neutre" entre individus. On suppose dans ce cas que l'on connaît la matrice de covariance des génotypes individuels, Σ, qui représente la structure "neutre" entre individus. On peut alors écrire un modèle de régression linéaire avec un bruit corrélé. Le modèle s'écrit de la manière suivante :

$$G_{i\ell}^{(c)} = X_i B_\ell^T + \epsilon_{i\ell},$$

où $\epsilon_\ell = (\epsilon_{1\ell}, \ldots, \epsilon_{n\ell})$ est un résidu de loi $N(0, \sigma_\ell^2 \Sigma)$ et B_ℓ est le vecteur contenant les coefficients de régression pour le locus ℓ.

L'estimateur du vecteur des coefficients de régression est donné par la formule suivante :

$$\hat{B}_\ell = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} G_\ell^{(c)}.$$

L'estimation du vecteur des coefficients de régression au locus ℓ est sans biais ($E(\hat{B}_\ell) = B_\ell$) et de variance minimale ($cov(\hat{B}_\ell) = \hat{\sigma_\ell^2}.(X^T \Sigma^{-1} X)^{-1}$).

Modèles GEE Dans un cadre plus général, Carl and Kühn (2007) ont proposé les modèles GEE pour "Generalized Estimating Equations". Ces modèles permettent de prendre en compte à la fois un bruit corrélé et une fonction de lien afin d'étudier l'auto-corrélation spatiale dans la distribution des espèces. Le modèle GEE a ensuite été repris par Poncet et al. (2010). Poncet et al. (2010) ont étudié la plante *Arabis alpina* dans les alpes françaises et suisses. L'utilisation de GEE leur a permis de détecter 78 locus significativement corrélés à la température chez la plante *Arabis alpina*.

Le modèle de régression linéaire à bruit corrélé permet donc de corriger la régression linéaire lorsque l'on connaît la forme des facteurs de confusion à travers une matrice de covariance entre individus, Σ. Toutefois, l'estimation ou la modélisation de cette matrice de covariance représente un vrai défi, généralement peu discuté par les utilisateurs de ces méthodes.

1.2.5.2 Modèle du logiciel `BAYENV`

L'un des modèles récemment utilisé en génétique des populations pour réaliser une étude d'association écologique en utilisant des méthodes corrélatives a été développé par Coop et al. (2010); Günther and Coop (2013). Le logiciel, du nom de `BAYENV`, propose d'estimer la matrice de covariance entre les populations grâce à un ensemble de SNPs de contrôle, putativement neutres. Nous décrivons ce modèle ci-dessous.

On considère un ensemble de locus bialléliques putativement neutres où l'on note $p_{j\ell}$ la fréquence d'allèle au locus ℓ dans la population j (n_j allèles ont été échantillonnés) et $k_{j\ell}$ le nombre d'allèles dérivés égaux à 1. Coop et al. (2010) supposent que $k_{j\ell}$ suit une loi binomiale de paramètre $p_{j\ell}$ et n_j,

$$p(k_{j\ell}|p_{j\ell}, n_j) \propto p_{j\ell}^{k_{j\ell}}(1-p_{j\ell})^{n_j-k_{j\ell}}.$$

Suivant le modèle de Nicholson et al. (2002), les fréquences d'allèles, $p_{j\ell}$, suivent une loi normale centrée autour de la fréquence ancestrale au locus ℓ, f_ℓ, tronquée dans l'intervalle $[0,1]$ et de variance proportionnelle à $f_\ell(1-f_\ell)$

$$p(p_\ell|\Omega, f_\ell) \sim N_{[0,1]}(f_\ell, f_\ell(1-f_\ell)\Omega),$$

où Ω représente la matrice de covariance des fréquences d'allèles entre populations, $p_\ell = (p_{j\ell})_j$ est le vecteur de fréquences d'allèles dans les populations au locus ℓ, et $N_{[0,1]}$ est la loi normale tronquée dans l'intervalle $[0,1]$.

Coop et al. (2010) supposent une loi a priori de la famille Inverse Wishart pour la matrice de covariance Ω. Ils calculent la distribution jointe de tous les paramètres et obtiennent un estimateur a posteriori de Ω en utilisant un ensemble de SNPs putativement neutres. Coop et al. (2010) testent alors la corrélation entre l'indicateur environnemental, X, et le vecteur de fréquences d'allèles dans les populations p_ℓ au locus ℓ. Ainsi, ils régressent les fréquences d'allèles par l'indicateur environnemental à travers un modèle de régression linéaire gaussien à bruit corrélé en utilisant $\hat{\Omega}$, l'estimateur de Ω, pour modéliser la structure entre les populations. Le vecteur des fréquences d'allèles au locus ℓ suit la loi suivante :

$$p(p_\ell|\hat{\Omega}, f_\ell, B_\ell, X) \sim N_{[0,1]}(f_\ell + XB_\ell^T, f_\ell(1-f_\ell)\hat{\Omega}),$$

où B_ℓ est le vecteur des coefficients de régression au locus ℓ.

Coop et al. (2010); Günther and Coop (2013) proposent ensuite de calculer un facteur de Bayes ou de faire un test de corrélation pour chaque SNP afin de déterminer les SNPs significativement associés à l'indicateur environnemental, X.

L'intérêt du modèle du logiciel `BAYENV` est d'avoir étendu le concept de régression linéaire corrélé tout en estimant la matrice de covariance entre populations. Une limite du modèle est d'être fondé sur les populations et non sur les individus.

Une autre limite de `BAYENV` est la nécessité d'identifier un ensemble de locus neutres, avant de tester l'association avec les facteurs environnementaux. Tout d'abord, le fait d'identifier une telle liste a priori peut engendrer un manque de puissance des tests fondés sur l'estimation empirique de Ω. C'est une limite importante pour des jeux de données où tous les SNPs sont potentiellement adaptatifs. Par exemple, des SNPs issus de données d'expression sont souvent utilisés pour étudier l'adaptation locale chez des organismes non-modèles (Eckert et al., 2010). De plus, le fait d'avoir à choisir un sous ensemble de marqueurs réduit la taille des données et un tel jeu de données peut biaiser arbitrairement les tests statistiques si seulement un sous-ensemble des données est choisi pour représenter les marqueurs neutres. En effet, il est possible que des marqueurs putativement neutres soient liés à des marqueurs sous sélection par déséquilibre de liaison sur de grandes distances physiques à travers le génome (Thibert-Plante and Hendry, 2010).

En pratique, les utilisateurs ont tendance à utiliser l'ensemble de leurs données pour calculer la matrice de covariance, Ω. D'un point de vue statistique, cela correspond à utiliser à deux reprises les données pour estimer les paramètres. Cela peut produire des effets indésirables et introduire de la variabilité entre les analyses (Blair et al., 2014). En effet, si les SNPs corrélés à l'environnement ont une forte influence sur la structure des populations, le fait d'utiliser ces SNPs pour estimer la structure neutre va entraîner une correction trop forte et faire perdre beaucoup de puissance à la méthode `BAYENV` (Frichot et al., 2013).

1.2.5.3 Modèles mixtes

Une limite de la régression linéaire est le fait qu'elle ne prenne pas en compte les facteurs de confusion propres à la génétique des populations. La solution envisagée dans le paragraphe précédent est un modèle de régression linéaire à bruit corrélé dans lequel on considère que les facteurs de confusion sont la conséquence d'une structure dite "neutre" entre individus.

Toutefois, les facteurs de confusion ne sont probablement pas communs à tous les SNPs. Il a donc été proposé, alternativement à la régression à bruit corrélé, d'utiliser des

modèles mixtes. L'idée principale des modèles mixtes en génomique des populations est de modéliser la variable X par la matrice des génotypes G en considérant que le modèle est la somme d'un effet fixe modélisant la corrélation avec l'indicateur environnemental X et un effet aléatoire modélisant les facteurs de confusion. L'avantage de l'effet aléatoire par rapport à une matrice de covariance est de pouvoir modéliser la variabilité de chaque individu.

Le modèle peut s'écrire de la manière suivante :

pour tout $1 \leq k \leq d$

$$X_k = (X_{1k}, \ldots, X_{nk})^T = G^{(c)} B_k^T + u_k + \epsilon_k$$

où X_k est la kième variable environnementale, $\epsilon_k \sim N(0, \sigma_k^2 I_n)$, $u_k \sim N(0, \Omega)$ et B_k est le vecteur des coefficients de régression. La matrice de covariance Ω est à déterminer et représente la structure commune à tous les facteurs de confusion.

Modèles mixtes dans les études d'association entre phénotypes et génotypes. Les modèles mixtes ont largement été étudiés et appliqués lors des études d'association entre génotypes et phénotypes (Abney et al., 2002; Yu et al., 2006; Aulchenko et al., 2007; Kang et al., 2008, 2010; Zhang et al., 2010; Price et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012; Zhou et al., 2013). Le but des études d'association est de décrire les corrélations entre des marqueurs génétiques et des traits phénotypiques à l'échelle d'une population (Cardon and Bell, 2001). Des études ont été réalisées pour de nombreux traits. Par exemple, les maladies chroniques héréditaires, comme le diabète, constituent l'un des enjeux majeurs pour les études d'association.

Une des limites des modèles mixtes est le choix de la matrice de covariance de l'effet aléatoire permettant de corriger pour les facteurs de confusion. Le choix de cette matrice peut influencer grandement les résultats. Le choix de la matrice de parenté semble cohérent en études d'association entre génotypes et phénotypes mais il semble avoir des limites pour les études d'association écologique. En effet, le phénotype est un trait héritable et donc lié aux génotypes, alors que l'indicateur environnemental n'est pas héritable. Il est la représentation de facteurs exogènes englobant un ensemble de pressions environnementales.

1.2.5.4 Résumé de la problématique

La détection de l'adaptation locale au sein des génomes représente une problématique importante. L'une des approches les plus utilisées pour réaliser une étude d'association écologique est de déterminer les SNPs corrélés avec un indicateur environnemental. Or il est nécessaire de prendre en compte les facteurs de confusion propre à la structure des données génétiques. Ces effets peuvent être dus à la structure "neutre" entre les individus, mais aussi à des effets propres à la variation de chacun des SNPs. Afin de prendre en compte la structure dite "neutre", des modèles de régression à bruit corrélé ont été développés. Une des difficultés de ces modèles est d'estimer la structure neutre entre individus sans biais. Afin de prendre en compte les variations propres à chacun des SNPs, les modèles mixtes ont ensuite été proposés. Toutefois ces effets sont structurés par une même matrice de covariance, qu'il faut choisir a priori. Il est donc nécessaire de développer une approche flexible afin de réaliser des études d'association écologique pour détecter de l'adaptation locale chez une espèce, permettant de corriger les biais dus aux nombreux facteurs de confusion impossibles à observer.

1.3 Objectifs de la thèse

Dans cette partie, nous présentons les objectifs de cette thèse tout en résumant les difficultés dont nous devons tenir compte et la façon dont nous avons choisi de les aborder.

Rendu possible par l'apparition de données génomiques et environnementales massives pour l'homme et pour de nombreuses espèces, l'objectif de la thèse est de développer des approches statistiques rigoureuses permettant d'analyser les différentes forces évolutives responsables de l'adaptation des espèces à leur environnement, incluant en particulier la réponse au climat, tout en prenant en compte la complexité des processus évolutifs.

Les problématiques pour répondre à cet objectif s'exercent selon deux axes principaux. La première direction est le passage à l'échelle ou comment appliquer des techniques existantes à des données dont le volume est multiplié par plusieurs ordres. La deuxième direction consiste à développer des méthodes innovantes prenant en compte la spécificité de données constituées de génomes entiers, afin de répondre à des questions posées par les écologues ou les généticiens des populations.

Pour répondre à ces problématiques, nous proposons d'utiliser des modèles à facteurs latents en génétique des populations pour modéliser les facteurs de confusion. Ce type de modèle a été largement étudié dans plusieurs domaines (Anderson and Gerbing, 1984).

Notamment, il a permis dans le domaine de l'apprentissage automatique d'étudier des données de très grandes tailles (plusieurs millions d'individus pour plusieurs millions de répétitions)(Koren et al., 2009). De plus, les modèles à facteurs sont des modèles très flexibles qui permettent de modéliser les spécificités des données de la génétique des populations (Duforet-Frebourg et al., 2014).

Enfin, le délivrable de cette thèse est un ensemble de logiciels multi-plateformes destinés à une communauté d'utilisateurs en écologie moléculaire, en épidémiologie et en génétique des populations permettant à la fois l'inférence de la structure des populations et la détection de gènes sous adaptation locale. Ces logiciels comportent une version en ligne de commande et une interface graphique permettant l'analyse de génomes entiers sur un ordinateur de bureau en un temps raisonnable. En particulier, nous avons développé une bibliothèque en langage R nommée LEA (*Landscape and Ecological Associations tests*) qui regroupe les différents logiciels de cette thèse utilisant la puissance des outils statistiques et les facilités propres au langage R. Cette bibliothèque a été conçue pour permettre aux généticiens des populations d'utiliser nos méthodes facilement afin de se concentrer sur la modélisation ainsi que sur la compréhension et l'interprétation des résultats.

Chapitre 2

Modèles à facteurs latents en génétique des populations

Dans cette partie, nous présentons les approches développées durant la thèse. Nous avons tout d'abord cherché à modéliser la structure de population à l'aide de modèles à facteurs latents. Dans un premier temps, nous résumons donc les différentes approches existantes fondées sur les modèles à facteurs latents pour étudier la structure génétique de population et nous expliquons notre contribution. Puis, dans un second temps, nous présentons une nouvelle approche pour corriger les tests d'association écologique prenant en compte les facteurs de confusion propres à la génétique des populations et s'appuyant sur un modèle à facteurs latents.

2.1 Étude de la structure génétique de population

Un ensemble d'individus est dit génétiquement structuré s'il existe des populations homogènes au sein de cet ensemble. De nombreux phénomènes peuvent entraîner l'existence de structure dans un groupe d'individus, tels que l'histoire ou les pressions environnementales. L'un des phénomènes les plus étudiés stipule simplement que plus les individus sont géographiquement isolés plus ils vont accumuler des différences génétiques (Malécot, 1948; Sokal and Oden, 1978; Slatkin and Arter, 1991). Wright (1943) a décrit ce phénomène qu'il a appelé *l'isolement par la distance*.

Un autre phénomène de première importance est *le métissage*. Lorsque les individus migrent massivement vers une population auparavant isolée de leur population d'origine, l'ensemble formé de la population hôte et du groupe de migrants est structuré

génétiquement. En effet, même si le groupe de migrants se reproduit aléatoirement avec les membres de la population hôte, une différence génétique va subsister entre les descendants des migrants et les autres pendant plusieurs générations.

Il est en général important de déterminer les populations d'origine des individus métissés. On cherche à déterminer la proportion du génome de chaque individu de l'échantillon issue de chacune des populations d'origine. Les groupes s'interprètent alors comme des populations parentales (ou ancestrales) qui se sont mélangées, donnant naissance à des individus métissés (ou hybrides).

Il existe deux types principaux d'approches pour étudier la structure génétique des populations : les approches de type "analyse en composantes principales" et les approches de type "structure". On parle d'approches de type "structure" en référence au logiciel développé par Pritchard et al. (2000a). Nous détaillons l'état de l'art dans ces deux approches et présentons nos contributions pour améliorer ces approches.

2.1.1 Analyse en composantes principales

L'un des outils les plus communs pour étudier la structure génétique des populations et plus généralement la structure de données multidimensionnelles est l'analyse en composantes principales (ACP). Dans cette section, nous rappelons le principe statistique de l'ACP, puis nous discutons le choix du nombre de composantes principales. Ensuite, nous présentons les applications de l'ACP en génétique des populations. Enfin, nous présentons notre contribution : un modèle d'ACP corrigeant pour l'isolement par la distance, appelé spFA (spatial Factor Analysis).

2.1.1.1 Principe de l'ACP

L'ACP est une méthode d'analyse de données multidimensionnelles. Le principe de l'ACP est de réduire la dimension d'un jeu de données composé d'un grand nombre de données corrélées (Jolliffe, 1986), tout en représentant au mieux la variation au sein des données. Pour cela, on détermine successivement des axes, appelés composantes principales (PCs), orthogonales les uns aux autres, tels que la projection des données sur chaque axe soit de variance maximale. Si l'on cherche à résumer notre matrice de données génotypiques en utilisant k PCs, on peut écrire la matrice de données sous sa décomposition en valeurs singulières

$$G = USV^T$$

où U est une matrice unitaire de taille $n \times k$, S est la matrice diagonale $k \times k$ contenant les k plus grandes valeurs propres de G, et V est une matrice unitaire de taille $L \times k$.

D'un point de vue statistique et géométrique, la décomposition en valeurs singulières en k composantes peut être vue comme la projection sur l'espace vectoriel de dimension k maximisant la variance expliquée des données. U est alors la matrice de projection des individus sur le sous-espace vectoriel de dimension k et V la matrice de projection des locus sur le sous-espace de dimension k.

2.1.1.2 Applications en génétique des populations

Patterson et al. (2006) ont décrit l'intérêt de l'étude de la structure de population avec l'ACP. Si le groupe d'individus étudiés est structuré en K populations, l'ACP va révéler $K-1$ axes de variation orthogonaux, correspondant à $K-1$ valeurs propres significatives au sens du test de Tracy-Widom. Nous définissons le test de Tracy-Widom dans la section suivante consacrée à la significativité de la structure détectée. De manière formelle, Patterson et al. (2006) ont montré que si l'on simule un nombre infini de SNPs selon le F-modèle, un modèle de divergence instantanée en K populations pour n individus développé par Balding and Nichols (1995), avec une faible divergence entre les K populations, alors la matrice de génotypes normée $G^{(n)}$ admet $K-1$ valeurs propres infinies, $n-K$ petites valeurs propres et une valeur propre nulle. Autrement dit, dans un modèle à K populations, l'ACP va détecter $k = K-1$ valeurs propres significatives au sens du test de Tracy-Widom.

2.1.1.3 Significativité de la structure détectée

Lorsque que l'on effectue une ACP, on cherche à déterminer l'ensemble des axes représentant significativement les différences entre les individus. Comme expliqué au paragraphe précédent, on peut alors chercher à déterminer $K-1$ axes significatifs pour différencier K populations. Or le nombre de populations à déterminer, K, est inconnu. Pour l'estimer, il existe deux méthodes principales : une méthode fondée sur le pourcentage de variance expliquée par chaque axe et une méthode fondée sur le test de Tracy-Widom, test de la nullité des valeurs propres de l'ACP.

Pourcentage de variance expliquée On sait que l'on peut associer à chaque composante principale le pourcentage de variance des données que cette composante explique. On peut alors choisir le nombre de composantes principales qui expliquent un certain

pourcentage de la variation des données comme étant le nombre de composantes significatives dans l'ACP. Toutefois, il peut être difficile de faire un choix avec des données génétiques réelles car il arrive que chaque composante explique une très faible proportion de la variance des données. De plus, lorsque le nombre de locus est grand devant le nombre d'individus, l'estimation du pourcentage de variance expliqué peut être biaisée (Lee et al., 2010).

Test de Tracy-Widom Johnstone (2001) a montré que si le nombre d'individus n et le nombre de locus L sont suffisamment grand, alors la plus grande valeur propre de la matrice des génotypes normée $G^{(n)}$ suit approximativement la loi de Tracy-Widom Tracy and Widom (1994). On peut alors réaliser un test pour chaque valeur propre, fondé sur la loi de Tracy-Widom, appelé test de Tracy-Widom, pour déterminer si cette valeur propre est significativement différente de 0 (Patterson et al., 2006). Cela nous permet de déterminer le nombre $k = K - 1$ de valeurs propres significatives et ainsi de regarder les différences entre les individus au sein du jeu de données avec les $k = K - 1$ premiers axes de l'ACP. En pratique, on observe, à l'aide de données simulées selon différents modèles de génétique des populations, que le test de Tracy-Widom a tendance à estimer un nombre de valeurs propres significatives plus grand que le nombre de populations simulées (Frichot et al., 2013).

2.1.1.4 Exemples d'étude de la structure des populations avec l'ACP

L'ACP s'est montrée utile pour l'analyse de la structure génétique dans de nombreuses études. Par exemple, Frichot et al. (2012) ont étudié la structure de population d'un échantillon de 418 individus répartis dans 27 populations asiatiques issus du Harvard Human Genome Diversity Project - Centre Étude Polymorphism Humain (Harvard 228 HGDP-CEPH) (ftp ://ftp.cephb.fr/hgdp_v3/). Sur la carte de la répartition des populations (Figure 2.1 A), on observe les populations d'Asie Centrale et les populations d'Asie de l'Est. La visualisation des deux premiers axes de l'ACP nous montre une répartition continue des populations allant de l'Asie Centrale à l'Asie de l'Est. Plus précisément on observe un effet "fer à cheval", signature de la présence de formes dues à l'isolement par la distance au sein des données (Figure 2.1 B) (Legendre and Gallagher, 2001; Diaconis et al., 2008).

L'influence du phénomène d'isolement par la distance sur l'ACP a été étudiée par Novembre and Stephens (2008). Plus précisément, Novembre et Stephens ont montré que le processus d'isolement par la distance tend à créer des formes spécifiques à l'ACP qui ne reflètent pas seulement la géographie. Pour cela, ils ont simulé des données génétiques

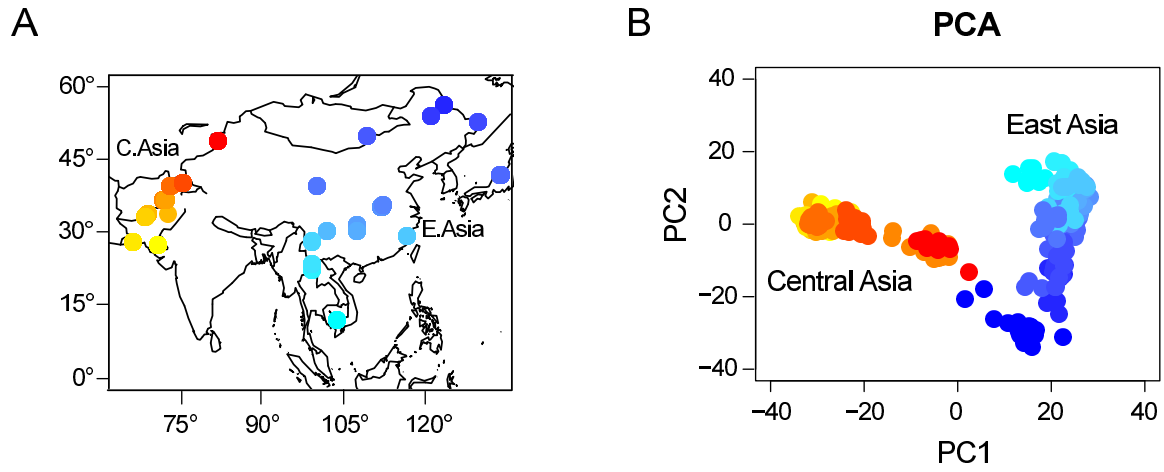

FIGURE 2.1: Analyse de la structure génétique de populations asiatiques. A) Carte représentant la position géographique des populations en Asie. Les populations d'Asie centrale sont représentées dans des couleurs allant du jaune au rouge. Les populations d'Asie de l'est sont représentées avec des couleurs bleues. B) Visualisation des deux premiers axes d'une ACP. Les individus sont projetés dans le repère (PC1, PC2).

selon un modèle en îles unidimensionnel (Figure 2.2 A). On peut voir que, malgré un habitat linéaire, les composantes principales décrivent des sinusoïdes (Figure 2.2 B). Cela a pour conséquence de donner une structure très particulière à la visualisation des deux premiers axes de l'ACP qui ne reflète pas seulement la géographie de l'habitat simulé (Figure 2.2 C).

D'autres méthodes, fondées sur la même idée que l'analyse en composantes principales, présentent le même type d'effet que l'ACP. On peut, par exemple, citer l'analyse principale en coordonnées (principal coordinates analysis (PCoA) aussi appelée multidimensional scaling (MDS) Pearson 1901, Kruskal 1978 ou sammon mapping Sammon 1969). Cette méthode consiste à déterminer les similarités entre les individus à partir d'une distance donnée. Dans le cadre de la génétique, on peut décider de choisir une distance génétique appropriée. Si l'on choisit la distance associée à la distance euclidienne, PCoA est similaire à l'ACP.

2.1.1.5 Modèles à facteurs latents

L'ACP fait partie d'une plus grande famille de méthodes appelées les modèles factoriels. L'idée principale des modèles à facteurs est de décomposer la matrice de génotypes en un produit de matrices ayant des propriétés spécifiques. Par exemple, l'ACP factorise une matrice de génotypes en une matrice unitaire de rang k, une matrice diagonale contenant les k plus grandes valeurs propres et une autre matrice unitaire de rang k. Dans notre cas, on s'intéresse à la décomposition de la matrice de génotypes en deux matrices U et

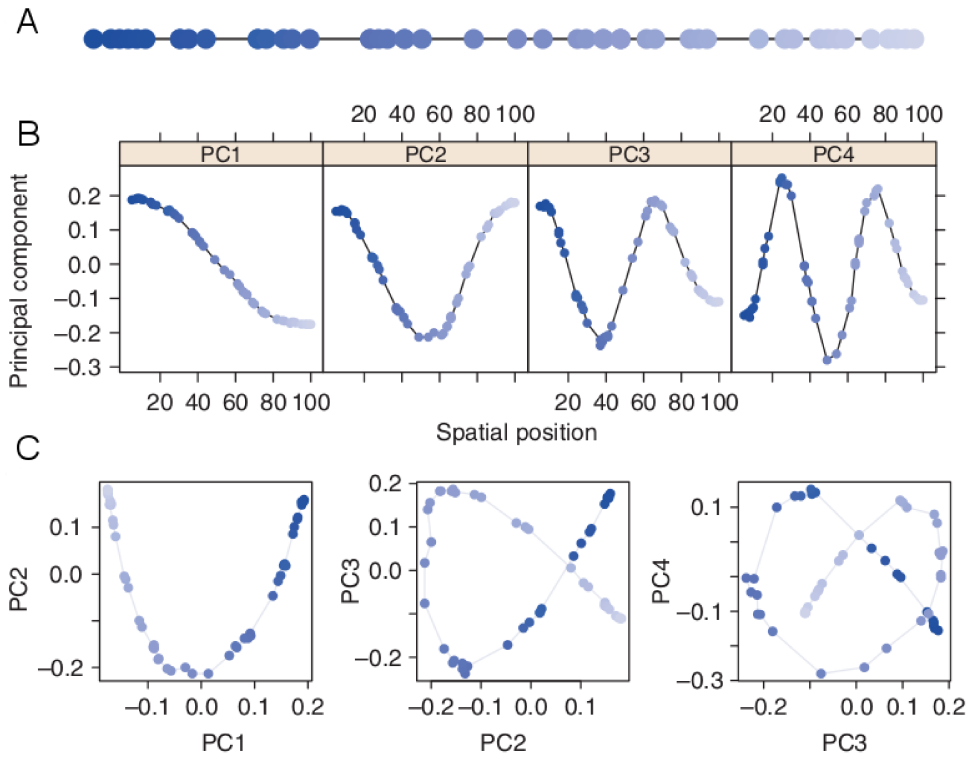

FIGURE 2.2: Résultats d'une ACP appliquée à des données issues d'un habitat unidimensionnel. A) Représentation de l'habitat, avec un cercle pour chaque population échantillonnée. B) Cartes des PCs avec un cercle pour chaque population. C) Visualisations deux à deux des axes de principaux de l'ACP (Novembre and Stephens, 2008).

V de rang faible k (avec k très petit devant n) sans propriété supplémentaire, comme illustré ci-dessous :

$$\begin{array}{c} \\ i \end{array} \begin{array}{c} \ell \\ \begin{pmatrix} \vdots \\ \dots \quad G_{i\ell} \quad \dots \\ \vdots \end{pmatrix} \end{array} = \begin{array}{c} k \\ \begin{pmatrix} \vdots \\ U_{ik} \\ \vdots \end{pmatrix} \end{array} \begin{array}{c} \ell \\ (\dots \quad V_{k\ell} \quad \dots) \end{array} \begin{array}{c} \\ k \end{array}$$

On parle alors de factorisation de matrice. L'avantage de la factorisation de matrice, modélisation plus flexible que celle de l'ACP, est de pouvoir considérer un cadre bayésien à la décomposition en facteurs. On peut alors considérer différentes lois a priori sur les facteurs U et V pour construire différents modèles, tels que l'ACP probabiliste (Tipping and Bishop, 1999), la factorisation de matrice bayésienne (Salakhutdinov and Mnih, 2008), l'analyse factorielle creuse (Engelhardt and Stephens, 2010), ou la factorization de matrice par maximisation de marge (MMMF, Srebro et al. 2004) l afin de mieux modéliser la variation des données. Cela permet aussi de considérer différents algorithmes pour estimer les facteurs (Maximum A Posteriori, Expectation-Maximization, Tipping and Bishop

1999, échantillonnage de Gibbs, Salakhutdinov and Mnih 2008, algorithme variationnel bayésien, Seeger and Bouchard 2012, Nakajima et al. 2013).

2.1.1.6 Notre contribution : spatial Factor Analysis

Isolement par la distance Dans le cadre de cette thèse, nous nous sommes intéressés au concept de l'isolement par la distance dans l'ACP. Le concept a été introduit par Wright pour décrire l'accumulation de différences génétiques locales au sein d'un espace de dispersion restreint (Wright, 1943). La théorie prédit que si une espèce est répartie de manière continue dans l'espace et si elle possède une faible dispersion géographique, alors la différenciation génétique croit avec la distance géographique (Malécot, 1948; Kimura and Weiss, 1964). L'isolement par la distance peut être décrit par l'*autocorrélation spatiale*, une mesure du degré de dépendance entre individus au sein d'un espace géographique (Rousset, 1997). L'autocorrélation spatiale peut poser des difficultés pour l'analyse des données de génétique des populations.

Modèle de spFA Nous avons proposé de corriger l'ACP pour les effets non désirables générés par l'isolement par la distance. Nous avons donc proposé une modélisation de l'ACP dont le bruit prend en compte la distance entre les individus. Pour cela, nous considérons un modèle à facteurs avec un bruit autocorrélé. De manière plus formelle, le modèle s'écrit de la manière suivante

$$G = UV^T + \epsilon,$$

où chaque colonne ϵ_ℓ de ϵ suit une loi normale centrée $N(0, \Sigma_\theta)$ de matrice de covariance Σ_θ avec

$$\Sigma_\theta(i,j) = e^{-\frac{d(i,j)}{\theta}}, \quad i,j = 1 \dots n,$$

où $d(i,j)$ est la distance entre l'individu i et l'individu j et θ est un paramètre d'échelle (Figure 2.3). Le paramètre θ est un paramètre d'échelle mesuré en unité de distance moyenne entre paires de positions géographiques, $\bar{d}$. Afin de pouvoir interpréter les facteurs U, on ajoute la contrainte que la matrice de covariance de U doit être diagonale et que V doit être une matrice unitaire. De plus, U et V sont de rang K, petit devant le nombre d'individus.

L'idée principale est de modéliser l'autocorrélation entre les individus à travers le bruit afin de séparer la partie de la variation génétique dûe à l'autocorrélation du reste de la variation génétique. Pour cela, nous modélisons la matrice de covariance du bruit comme

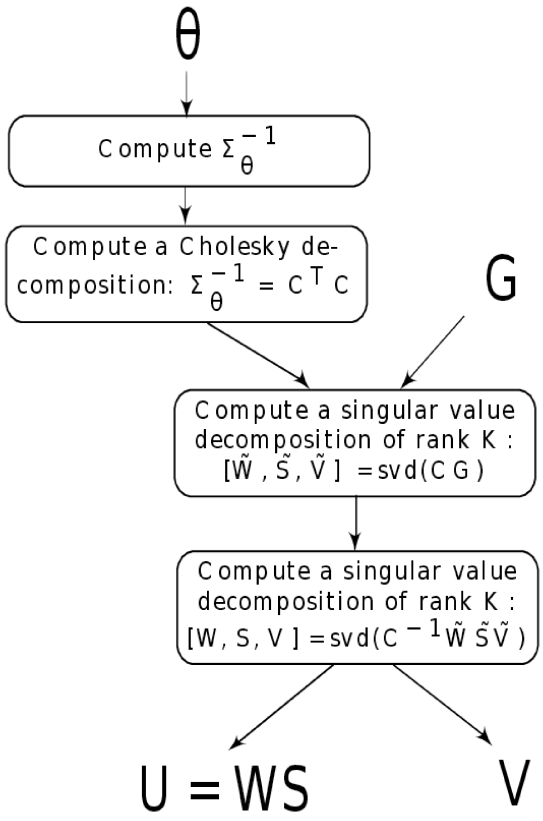

FIGURE 2.3: Algorithme de spFA. Pour une matrice de génotypes, G avec des coordonnées géographiques (X_i) et un paramètre d'échelle $\theta > 0$, la figure décrit les étapes de spFA.

décrit ci-dessus car on peut approximer les vecteurs propres de cette matrice de covariance avec les colonnes d'une transformation en cosinus discret, correspondant aux sinusoïdes que l'on observe dans les graphiques de l'ACP (Ahmed et al., 1974; Diaconis et al., 2008). Nous avons appelé ce modèle *Analyse Factorielle Spatiale* (spFA).

Résultats résumés Dans un premier temps, nous avons simulé des données génétiques selon un modèle en îles unidimensionnel suivant Novembre and Stephens (2008) (Figure 2.4 A). Puis nous avons appliqué spFA avec différentes valeurs du paramètre d'échelle θ. On observe qu'en augmentant graduellement la valeur du paramètre d'échelle, on fait disparaître les sinusoïdes de la carte de PC3, puis de PC2, puis de PC1, signature de l'isolement par la distance (Figure 2.4 B, C, D). spFA est donc capable de corriger à différentes échelles les effets dus à l'isolement par la distance. Nous proposons d'utiliser comme critère pour choisir le paramètre d'échelle, la statistique Λ de Wilks. Toutefois, la statistique Λ de Wilks présente des limites puisqu'il est nécessaire d'assigner au préalable les individus en K groupe pour la calculer.

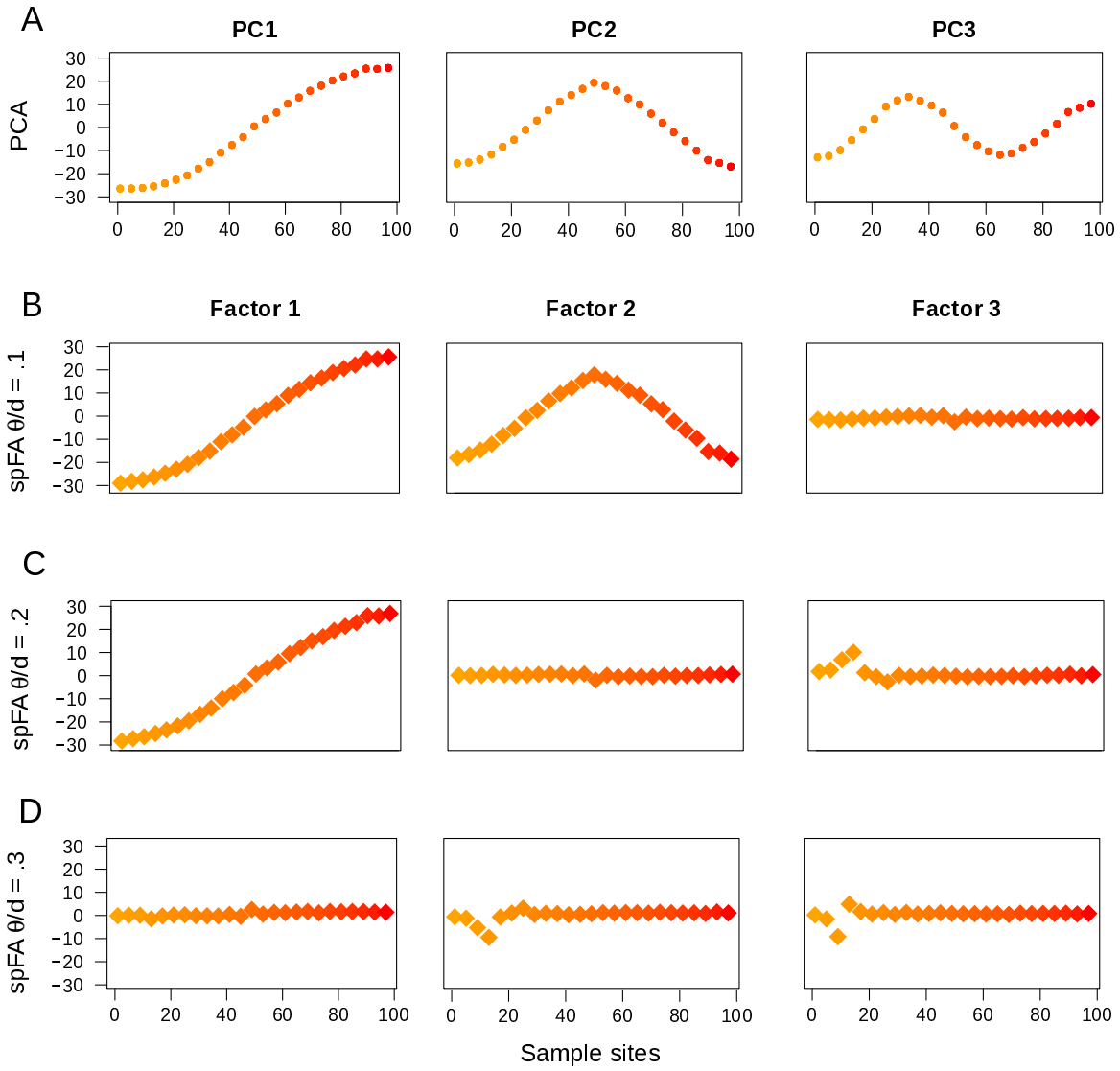

FIGURE 2.4: Carte de PCs pour l'ACP (A) et spFA pour différentes échelles de θ (B-D).

Nous avons étudié la structure d'un échantillon de 418 humains répartis dans 27 populations asiatiques, précédemment étudié avec l'ACP (section 2.1.1.4). Pour rappel, la visualisation des deux premiers axes de l'ACP sur ce jeu de données nous montre une répartition continue des populations allant de l'Asie Centrale à l'Asie de l'Est. On observe une répartition en forme de fer à cheval, signature de la présence d'isolement par la distance au sein des données (Figure 2.5 A). Contrairement aux deux premières composantes de l'ACP, les deux premiers facteurs de spFA nous montrent une discontinuité majeure séparant les populations d'Asie Centrale et les populations d'Asie de l'Est. De plus, les individus des populations Uyghur et Hazara sont alignés entre les deux groupes. Cela suggère que ces deux populations sont issues du métissage des populations ancestrales d'Asie Centrale et d'Asie de l'Est (Figure 2.5 B). Ainsi, spFA permet de corriger l'ACP

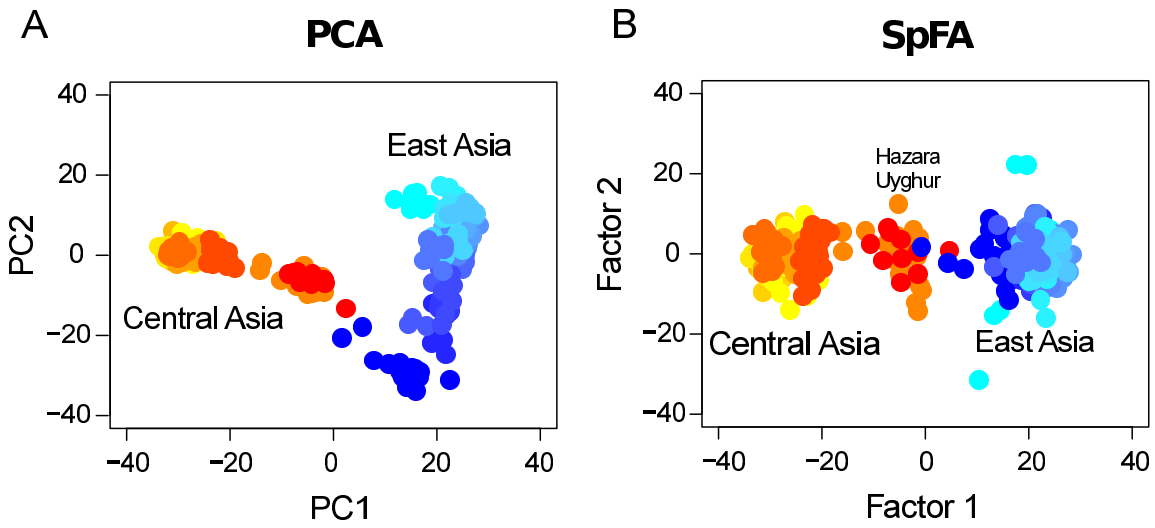

FIGURE 2.5: Analyse de la structure génétique de populations asiatiques. A) Visualisation des deux premiers axes de l'ACP. B) Visualisation des deux premiers facteurs de spFA.

pour le biais dû à l'isolement par la distance et permet de mieux interpréter les cartes des facteurs latents. Le détail des résultats concernant spFA est disponible au chapitre 3.

2.1.1.7 Conclusion sur les méthodes de type ACP

Dans cette partie, nous avons présenté une première façon de modéliser la variation génétique à travers l'analyse en composantes principales et de manière plus générale avec la factorisation de matrice. Nous avons vu que l'analyse en composantes principales permet de résumer la variation génétique des données grâce à un nombre réduit d'axes principaux. Ces axes permettent de visualiser les différences génétiques entre les individus et ainsi d'observer la structure génétique des populations. Puis, nous avons présenté spFA, une extension de l'ACP prenant explicitement en compte la répartition géographique des individus afin de tenir compte de l'isolement par la distance. Nous avons montré que spFA est capable de produire des facteurs principaux corrigés pour les formes sinusoïdales que l'on observe en cas d'isolement par la distance. Cela facilite l'interprétation des facteurs principaux et par conséquent, facilite l'étude de la structure génétique des populations.

2.1.2 Approches de type "structure"

L'autre approche, très utilisée pour étudier la structure génétique de population est une approche que l'on dit de type "structure". Dans ce type d'approche, on suppose que

l'échantillon est structuré en K populations ancestrales. Le nombre de populations K est fixe et un paramètre important à déterminer. On cherche alors à estimer pour chaque individu i, la proportion du génome de cet individu qui provient de chaque population ancestrale. On parle de modèle de métissage et la proportion de génome ancestral pour chaque individu, appelée *coefficient de métissage*, est notée Q_{ik} pour la population ancestrale k et l'individu i. De même, on cherche à déterminer pour chaque locus, sa fréquence d'allèle dans chacune des populations ancestrales. On parle alors de *fréquence ancestrale*, notée $F_{\ell k}$ au locus ℓ dans la population ancestrale k. Dans cette partie, nous présentons les approches principales utilisées pour estimer les coefficients de métissage ainsi que les hypothèses sous-jacentes aux modèles utilisés.

2.1.2.1 Le logiciel `STRUCTURE`

Le modèle de `STRUCTURE` (Pritchard et al., 2000a) fut le premier modèle bayésien pour l'analyse non supervisée de la structure génétique d'un échantillon d'individus. Ce logiciel calcule la proportion de métissage du génome de chaque individu. Le modèle du logiciel `STRUCTURE` a subi de nombreuses évolutions (Falush et al., 2003, 2007; Hubisz et al., 2009). Falush et al. (2003) ont proposé une extension du modèle qui autorise la liaison entre les locus. En effet, le modèle prend en compte les corrélations entre les locus liés qui proviennent de populations métissées. Nous nous limitons à la description du modèle avec métissage pour des individus diploides (Pritchard et al., 2000a). L'hypothèse faite est que chaque génotype, $G_{i\ell}$, observé pour l'individu i au locus ℓ pour l'allèle a provient d'un groupe inconnu, $Z^a_{i\ell}$. L'objectif de la modélisation est d'estimer la matrice $Q = (Q_{ik})_{ik}$, regroupant les coefficients de métissage.

Pour cela, le logiciel `STRUCTURE` simule, grâce à un algorithme d'échantillonnage de Gibbs, la loi a posteriori de (Z, Q, F) donnée par

$$p(Z, F, Q|G) \propto p(G|Z, F)p(Z|Q)p(Q)p(F)$$

où

$$P(G_{i\ell} = j|Z^j_{i\ell} = k, F) \propto F^j_{k\ell}(1 - F_{k\ell})^{2-j}, \quad j = 0, 1, 2,$$

et

$$p(Z^j_{i\ell} = k|Q) = q_{ik}.$$

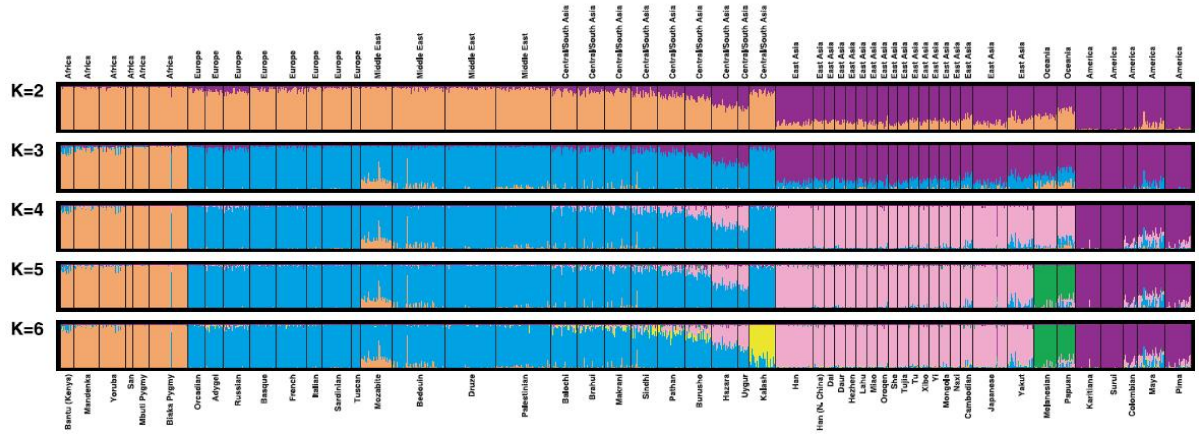

FIGURE 2.6: Diagramme en barres des résultats du logiciel `STRUCTURE` d'analyse d'un jeu de données de microsattelites du HGDP (Rosenberg et al., 2002).

Conformément aux modèles classiques de la génétique des populations, les fréquences alléliques suivent une loi de Dirichlet

$$F_{k\ell} \sim D(\lambda_1, \ldots, \lambda_{J_\ell}),$$

où J_ℓ représente le nombre d'allèles au locus ℓ et les λ_j sont fixés à 1 par défaut. La loi de Dirichlet est aussi utilisée pour modéliser les coefficients de métissage

$$Q_i \sim D(\alpha_{i_1}, \ldots, \alpha_{i_K}),$$

où Q_i est le vecteur des coefficients de métissage pour l'individu i et où α_{i_k} est proportionnel au coefficient de métissage moyen sur l'ensemble des individus. Le coefficient α_{i_k} peut être fixé ou estimé au cours de l'algorithme grâce à un algorithme de Métropolis-Hasting.

Pour définir la log-vraisemblance du modèle du logiciel `STRUCTURE`, on suppose que les groupes ancestraux sont à l'équilibre de Hardy-Weinberg, et que les locus sont en équilibre de liaison. L'hypothèse d'Hardy-Weinberg suppose que la fréquence de chaque génotype $(0, 1, 2)$ est issue d'un tirage de la loi binomiale avec deux réalisations et de probabilité la fréquence de l'allèle muté. Cela permet de construire la loi de chaque allèle à chaque locus pour tout génotype à partir de tirages indépendants (Pritchard et al., 2000a). Cette log-vraisemblance s'écrit de la manière suivante :

$$L(Q, F) = \sum_i \sum_\ell (G_{i\ell} \log(\sum_k Q_{ik} F_{k\ell}) + (2 - G_{i\ell}) \log(\sum_k Q_{ik}(1 - F_{k\ell}))).$$

Tout au long de l'évolution du logiciel, des modifications pour améliorer ses performances ont été proposées. La majorité d'entre elles repose sur l'utilisation d'algorithmes EM

(Expectation-Maximisation) (Tang et al., 2005; Chen et al., 2006; Wu et al., 2006). Plus récemment, Raj et al. (2014) proposent d'estimer les paramètres en utilisant une approximation variationnelle bayésienne. Cette dernière version est implantée dans le programme `fastSTRUCTURE`.

Le logiciel `STRUCTURE` est l'un des logiciels les plus utilisés en génétique des populations. Parmi les exemples les plus connus, Rosenberg et al. (2002) ont appliqué le logiciel `STRUCTURE` à 1 056 individus issus de 52 populations humaines pour lesquels 377 marqueurs microsatellites ont été génotypés (données HGDP). La matrice Q est représentée forme de diagramme en barres où chaque individu est représenté par une fine barre verticale partitionnée en K segments correspondant aux coefficients de métissage estimés (Figure 2.6). Avec un modèle à $K = 5$ populations, les auteurs retrouvent les 5 régions géographiques majeures du globe : Afrique, Eurasie, Asie de l'Est, Australie, Amériques.

Choix du nombre de populations ancestrales Le choix du nombre de populations ancestrales décrivant au mieux des données est un problème difficile. Ce choix peut varier selon l'échelle, l'échantillonnage des individus et bien d'autres paramètres du modèle. Pour choisir le nombre de groupes, K, Pritchard et al. (2000a) ont proposé de calculer un critère de déviance bayésienne de la forme

$$L(K) = \mu + \frac{\sigma^2}{4},$$

où μ et σ^2 sont la moyenne et la variance de la déviance sous la loi a posteriori. En faisant l'hypothèse que la déviance est gaussienne, Pritchard et al. (2000a) ont montré que la grandeur $-2 \log P(K)$ est proportionnelle au critère de déviance, $L(K)$, où $P(K)$ est la probabilité a posteriori qu'il y ait K groupes. Ainsi, le critère de déviance peut être interprété, à une constante près, comme le logarithme de la probabilité a posteriori qu'il y ait K groupes. En effectuant des simulations, Evanno et al. (2005) ont montré que le critère $L(K)$ atteint un plateau ou continue à augmenter après que la vraie valeur de K ait été atteinte. Pour éviter ce problème, Evanno et al. (2005) ont proposé un critère ad-hoc appelé ΔK qui approche la dérivée au second ordre de $L(K)$, et qui, une fois maximisé, permet de trouver une bonne valeur pour K. Un troisième critère utilisé pour choisir K est le critère d'information de déviance qui s'écrit comme la somme d'un terme d'ajustement et d'un terme de complexité (DIC, Spiegelhalter et al. 2002).

Malgré la sophistication de tous ces critères, il faut se garder de trop interpréter la valeur optimale de K en la considérant comme le nombre de populations dont sont issues les individus. En effet, ce nombre est en particulier sensible au nombre de marqueurs moléculaires utilisés et à la stratégie d'échantillonnage utilisée (Jay et al., 2012).

2.1.2.2 Le logiciel ADMIXTURE

Avec l'émergence du domaine de la génomique des populations et des jeux de données comportant des milliers voire des millions de SNPs, l'algorithme bayésien du logiciel STRUCTURE montre une limite de capacité de calcul en un temps raisonnable. Alexander et al. (2009) proposent une méthode plus rapide implémentée dans le logiciel ADMIXTURE. Cette méthode optimise la vraisemblance du modèle du logiciel STRUCTURE en utilisant une méthode de descente par bloc couplée avec une accélération de la convergence (méthode quasi-Newton).

Pour corriger la tendance des critères du logiciel STRUCTURE à surestimer le nombre de populations ancestrales, Alexander and Lange (2011) proposent un critère prédictif pour estimer le nombre de populations ancestrales : *la validation croisée*. Le principe du critère de validation croisée est d'effectuer l'apprentissage des paramètres sur une partie des données puis d'évaluer la probabilité de l'autre partie des données conditionnellement aux valeurs des paramètres estimés dans la première phase.

Plus formellement, le modèle du logiciel ADMIXTURE prédit les données grâce à la formule suivante :

$$e_{i\ell} = 2 \sum_k Q_{ik}^{ADM} F_{k\ell}^{ADM}$$

où $e_{i\ell}$ est l'espérance du génotype masqué au locus ℓ pour l'individu i. Puis l'erreur de prédiction est estimée en moyennant le carré de la déviance des résidus pour le modèle binomial (Alexander and Lange, 2011),

$$d(G_{i\ell}, e_{i\ell}) = G_{i\ell} \log(\frac{G_{i\ell}}{e_{i\ell}}) + (2 - G_{i\ell}) \log(\frac{2 - G_{i\ell}}{2 - e_{i\ell}}).$$

Le logiciel ADMIXTURE a permis, d'une part, d'estimer les coefficients de métissage avec une vitesse de l'ordre de 100 fois plus grande que le logiciel STRUCTURE. Cela a permis d'analyser des jeux de données contenant des milliers de SNPs. D'autre part, le logiciel ADMIXTURE a proposé un critère prédictif afin d'estimer le nombre de populations ancestrales. Le critère de validation croisée d'ADMIXTURE permet de choisir le nombre de populations ancestrales qui prédit au mieux les données génétiques observées.

Toutefois, le logiciel ADMIXTURE comporte des limites. D'une part, minimisant la vraisemblance du modèle du logiciel STRUCTURE, le modèle du logiciel ADMIXTURE comporte les mêmes hypothèses de modélisation que le modèle du logiciel STRUCTURE. En particulier, leur modèle commun suppose que les SNPs sont sous équilibre d'Hardy-Weinberg. Par conséquent, il n'est pas possible d'analyser les données génétiques issues d'échantillons

d'une espèce consanguine. Parmi les exemples connus d'espèces consanguines, on peut citer la plante modèle *Arabidopsis thaliana* (Atwell et al., 2010). D'autre part, le logiciel `ADMIXTURE` montre une limite de capacité de calcul avec l'arrivée de données génomiques comportant des millions de SNPs.

2.1.2.3 le logiciel `SFA`

Engelhardt and Stephens (2010) proposent d'unifier les analyses de type "structure" et l'ACP en utilisant un modèle d'analyse factorielle parcimonieuse (SFA, Sparse Factor Analysis). Leur méthode propose de décomposer la matrice de génotypes en deux matrices de tailles $n \times K$ et $K \times L$ comme le fait l'ACP et plus généralement la factorisation de matrice, tout en groupant les individus. De plus, le logiciel `SFA` est capable d'analyser des jeux de données issus du génotypage de génomes complets en un temps raisonnable.

Toutefois, une difficulté principale de `SFA` est que les coefficients d'ascendance issus du modèle, comme les composantes principales d'une ACP, ne satisfont pas les conditions requises pour être des coefficients de métissage (positivité et somme des proportions à 1) (Engelhardt and Stephens, 2010).

2.1.2.4 Conclusion sur les approches de type "structure"

Dans cette partie, nous avons présenté une autre façon de modéliser la structure génétique, à travers l'estimation de coefficients de métissage. Le modèle de Engelhardt and Stephens (2010) met en avant une similarité de modélisation entre les approches de type "structure" et l'ACP qui tend à faire penser que l'on peut unifier les deux approches. Un certain nombre de difficultés ont pu être mises en avant par les modèles de l'état de l'art. Tout d'abord, l'hypothèse d'Hardy-Weinberg empêche l'estimation des coefficients de métissage pour des espèces consanguines. Ensuite, les modèles actuels ne sont pas capables d'analyser des jeux de données à haute résolution comportant des millions de SNPs. Hors on retrouve, à l'heure actuelle, ce type de données massives dans de nombreux projets.

2.1.3 Notre contribution : sparse Non-negative Matrix Factorization (sNMF)

Dans le cadre de cette thèse, nous nous sommes intéressés à l'estimation de coefficients de métissage pour des jeux de données à haute résolution comportant des millions de SNPs.

Afin de construire un modèle flexible et capable de s'abstraire de l'hypothèse d'Hardy-Weinberg, nous avons repris l'approche amorcée par Engelhardt and Stephens (2010). Ainsi, nous proposons une nouvelle méthode d'estimation des coefficients de métissage fondée sur la décomposition en facteurs de la matrice de génotypes. Cette approche s'appuie sur une *factorisation de matrice non négative parcimonieuse* (sNMF, Kim and Park 2007). Dans cette partie, nous résumons notre travail en expliquant notre modélisation, décrivons une comparaison avec le logiciel `ADMIXTURE` et des analyses de données. Ce travail est présenté en détails dans le chapitre 4.

2.1.3.1 Modèle du logiciel `sNMF`

Nous proposons de reformuler la matrice de génotypes G afin d'utiliser un codage binaire. Supposons, pour faciliter l'explication, que nos locus sont bialléliques, des SNPs. Nous considérons 3 colonnes pour coder chaque locus. Si l'allèle de ce locus est $G_{i\ell} = 0$, on le code 100. Si $G_{i\ell} = 1$, on le code 010 et si $G_{i\ell} = 2$, on le code 001. La matrice, représentant les données génétiques en codage binaire est de taille $n \times 3L$, est notée $G^{(b)}$. Nous utilisons un codage binaire afin de pouvoir estimer séparément la fréquence d'allèle ancestrale de chaque génotype. De ce fait, le logiciel `sNMF` ne nécessite pas que le jeu de données vérifie l'hypothèse d'Hardy-Weinberg contrairement aux logiciels `STRUCTURE` et `ADMIXTURE`. Hors, cette hypothèse n'est pas vérifiée si l'on analyse des échantillons issus d'espèces consanguines. En revanche, le logiciel `sNMF` permet d'analyser ce type de données.

De la même manière que Engelhardt and Stephens (2010), nous cherchons à décomposer $G^{(b)}$ en une matrice de coefficients de métissage Q (de taille $n \times K$) et une matrice de fréquences ancestrales de génotypes $F^{(b)}$ (de taille $K \times 3L$). On note $F^{(b)}_{k\ell}(j)$ la fréquence de l'allèle j au locus ℓ dans la population ancestrale k. Avec les contraintes associées, le modèle du logiciel `sNMF` s'écrit de la manière suivante

$$G^{(b)} = QF^{(b)} \qquad Q \geq 0, F^{(b)} \geq 0$$

où on a $\sum_{k=1}^{K} Q_{ik} = 1$ et $\sum_{j=0}^{2} F^{(b)}_{k\ell}(j) = 1$ pour chaque locus ℓ et chaque individu i.

Pour déterminer les paramètres de ce modèle, on utilise un algorithme des moindres carrés alternés. L'algorithme des moindres carrés alternés met à jour itérativement Q et $F^{(b)}$ en optimisant les fonctions de moindres carrés suivantes :

$$LS_1(F^{(b)}) = ||G^{(b)} - QF^{(b)}||^2_F, \quad F^{(b)} \geq 0$$

TABLE 2.1: **Jeux de données utilisés.**

Jeu de données	n	L	Référence
HGDP00778	934	78K	(Patterson et al., 2012)
HGDP00542	934	48.5K	–
HGDP00927	934	124K	–
HGDP00998	934	2.6K	–
HGDP01224	934	10.6K	–
HGDP-CEPH	1,043	660K	(Li et al., 2008)
1000 Genomes	1,092	2.2M	(1000 Genomes Project Consortium., 2012)
A. thaliana	168	216K	(Atwell et al., 2010)

$$\mathrm{LS}_2(Q) = \left\| \begin{pmatrix} {F^{(b)}}^T \\ \sqrt{\alpha}\ e_{1\times K} \end{pmatrix} Q - \begin{pmatrix} {G^{(b)}}^T \\ 0_{1\times n} \end{pmatrix} \right\|_F^2, \quad Q \geq 0. \tag{2.1}$$

Dans ces notations, $||.||_F$ est la norme de Frobenius, α est un paramètre de régularisation permettant une meilleure estimation des coefficients de métissage, $e_{1\times K}$ un vecteur contenant seulement des 1 et $0_{1\times n}$ un vecteur contenant seulement des 0. On minimise LS_1 et LS_2 alternativement en utilisant une méthode de pivot par blocs principaux, proposée par Kim and Park (2011).

2.1.3.2 Comparaison avec le logiciel `ADMIXTURE`

Pour comparer notre approche avec celle du logiciel `ADMIXTURE`, nous avons utilisé un ensemble de jeux de données issus de la littérature (Table 2.1). Cinq jeux de données proviennent du projet HGDP pour lesquels on a corrigé le biais de recrutement (Patterson et al., 2012). Les 5 jeux contiennent le même nombre d'individus mais les SNPs sont différents. Un 6ème jeu est le jeu complet du HGDP (Li et al., 2008). Le 7ème jeu de données provient du projet de génotypage humain 1000 Genomes (1000 Genomes Project Consortium., 2012). Le dernier jeu de données vient du génotypage de la plante modèle *Arabidopsis thaliana* en Europe (Atwell et al., 2010). Cette plante à la particularité d'être extrêmement consanguine. Elle ne vérifie donc pas l'hypothèse d'équilibre d'Hardy-Weinberg dont on cherche à s'affranchir.

Tout d'abord, nous avons comparé les estimations des coefficients de métissage obtenus par notre approche et l'approche du logiciel `ADMIXTURE` sur les différents jeux de données du projet HGDP en fonction du paramètre de régularisation. Pour chaque analyse avec

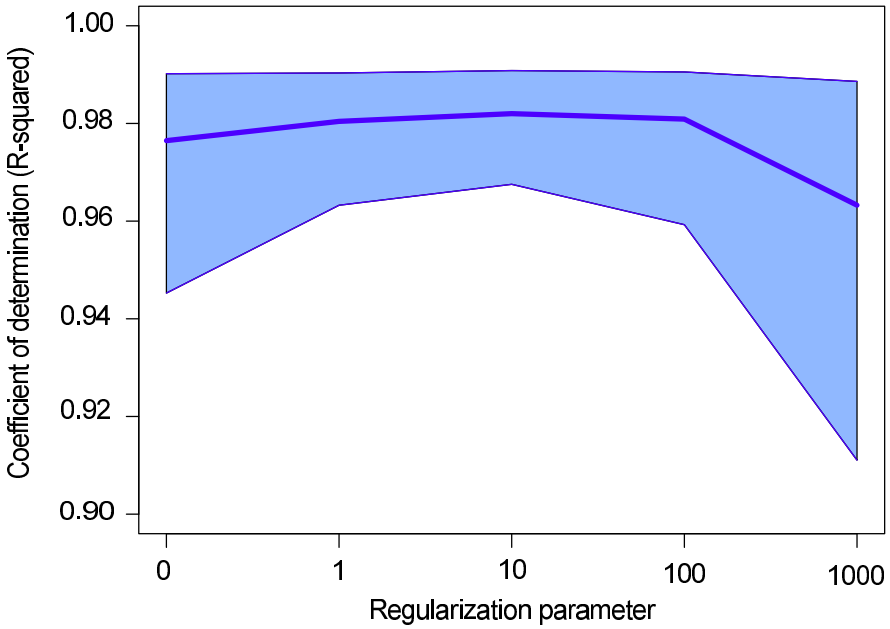

FIGURE 2.7: Graphique des coefficients de détermination entre les résultats du logiciel `sNMF` et du logiciel `ADMIXTURE` pour 5 jeux de données humaines en fonction du coefficient de régularisation.

le logiciel `ADMIXTURE`, nous avons calculé le coefficient de détermination maximal (R^2) parmi les résultats obtenus avec le logiciel `sNMF` ayant le même nombre de groupes (K). Pour K allant de 5 à 10, les coefficients de détermination restent supérieurs à 0.96 pour toutes les analyses (480 analyses, Figure 2.7). Ces résultats montrent que le logiciel `sNMF` permet d'obtenir des résultats très proches de ceux du logiciel `ADMIXTURE` pour les jeux du HGDP.

Pour le jeu de données Harvard HGDP panel HGDP00778 ($K = 7$), les analyses des deux programmes sont très proches ($R^2 = 0.99$, Figure 2.8 A), même si les résultats des deux logiciels sont sensible à l'effet de l'initialisation. Par exemple, lors d'analyses du jeu de données HGDP-CEPH, le logiciel `ADMIXTURE` identifie des groupes séparant les populations africaines de chasseurs-cueilleurs des autres populations, tandis que le logiciel `sNMF` identifie un unique groupe. De son côté, le logiciel `sNMF` sépare les populations du Moyen-Orient des populations européennes (Figure 2.8 B).

Puis nous avons comparé les temps de calcul entre le logiciel `ADMIXTURE` et le logiciel `sNMF` pour K variant entre 5 et 20, pour des jeux de données de différentes tailles (10.6K SNPs, 78K SNPs, 660K SNPs). On observe que le logiciel `sNMF` est 10 à 30 fois plus rapide que le logiciel `ADMIXTURE`. De plus, le temps de calcul évolue de manière quadratique par itération avec le nombre de populations ancestrales pour `ADMIXTURE`, tandis qu'il semble évoluer de manière linéaire avec K pour le logiciel `sNMF` (Figure 2.9).

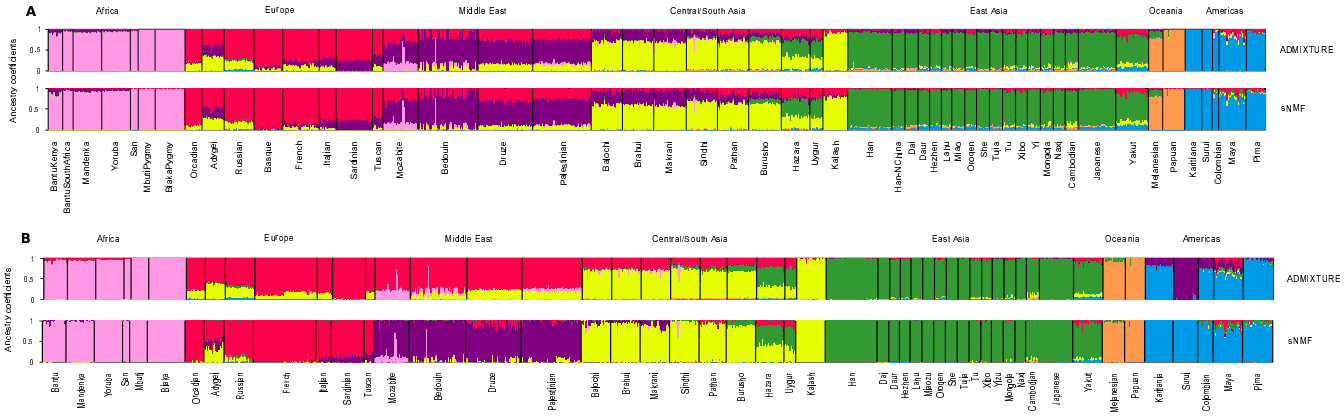

FIGURE 2.8: Exemple de représentation graphique des coefficients de métissage pour 2 jeux du HGDP. A) HGDP00778, 78K SNPs B) HGDP-CEPH (660K SNPs) pour $K = 7$ populations ancestrales.

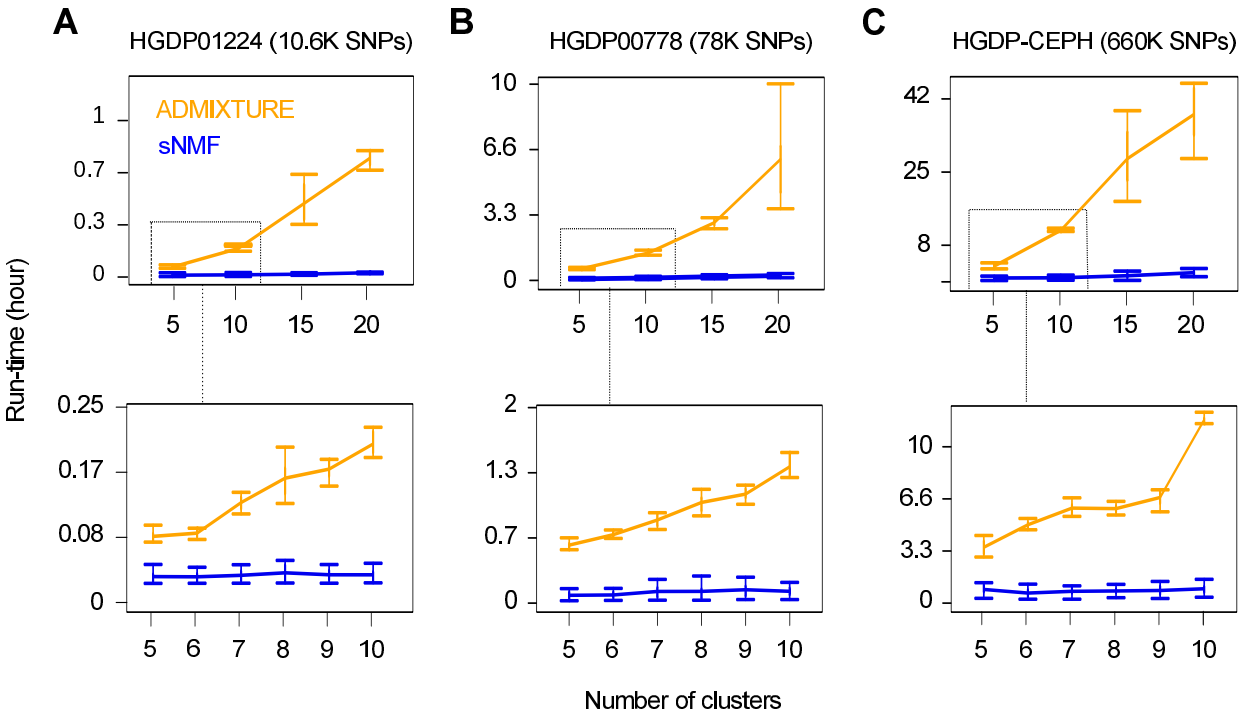

FIGURE 2.9: Temps de calcul du logiciel `sNMF` (bleu) et `ADMIXTURE` (orange) pour différentes valeurs de K et différents nombres de SNPs (10K SNPs, 78K SNPs, 660K SNPs).

2.1.3.3 Choix de nombre de populations ancestrales

De manière similaire au critère de validation croisée proposé par Alexander and Lange (2011), nous proposons un critère prédictif pour déterminer le nombre de populations

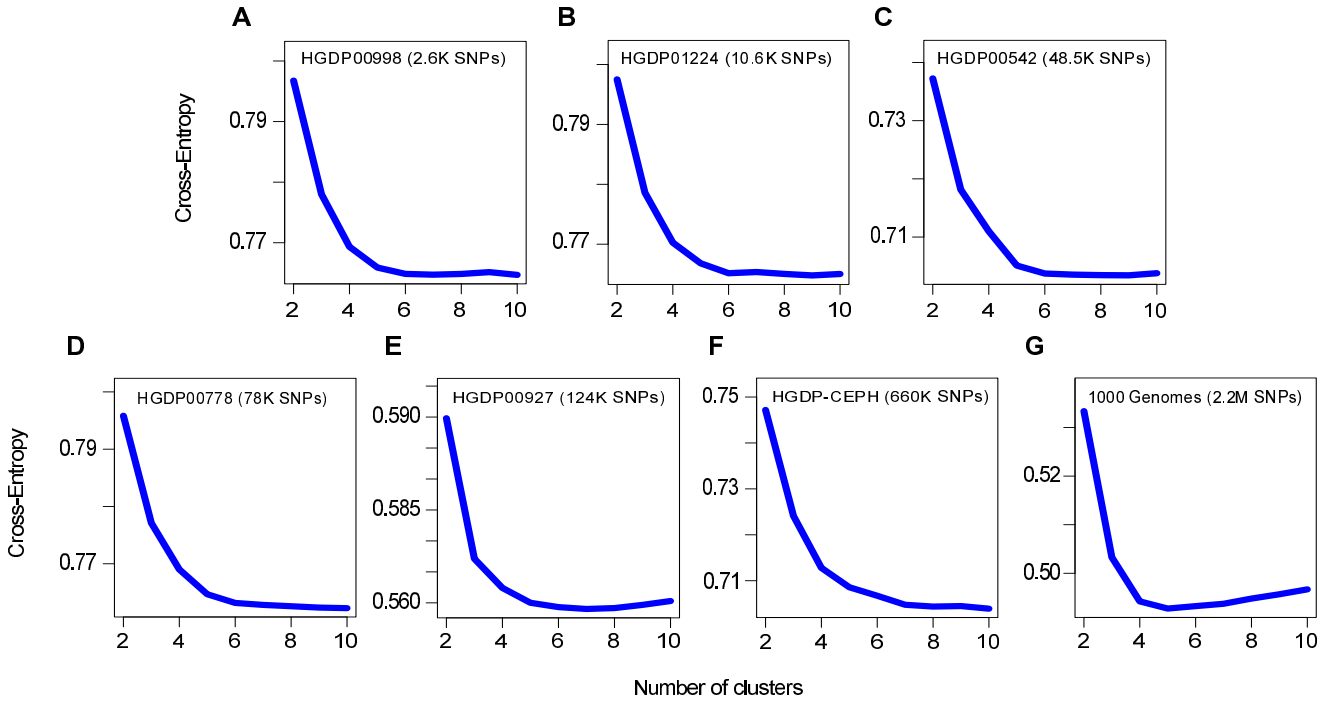

FIGURE 2.10: Critère d'entropie croisée pour tous les jeux de données humains en fonction de K, le nombre de populations ancestrales.

ancestrales à considérer. Ce critère est appelé le *critère d'entropie croisée*.

Pour calculer ce critère, nous enlevons un faible pourcentage des génotypes (5% par exemple). Nous appellerons ces génotypes *des génotypes masqués*. Puis, nous estimons les paramètres du modèle, Q et $F^{(c)}$. Nous calculons ensuite la probabilité de chaque génotype masqué, $G_{i\ell}$ pour l'individu i au locus ℓ de la manière suivante :

$$p_{i\ell}^{pred}(j) = \sum_{k=1}^{K} Q_{ik} F_{k\ell}^{(c)}(j), \quad j = 0, 1, 2.$$

Nous définissons le critère d'entropie croisée comme la moyenne sur les génotypes masqués de la grandeur $-\log p_{i\ell}^{pred}(G_{i\ell})$.

Nous avons calculé le critère d'entropie croisée pour l'ensemble des jeux de données humaines (Table 2.1). Un plus petit critère d'entropie croisée correspond à une meilleure prédiction du génotype masqué. On observe que la meilleure valeur de K dépend du jeu de données utilisé (Figure 2.10). Les résultats sont proches de ceux obtenus lors d'études précédentes pour des données humaines (Rosenberg et al., 2002; Li et al., 2008).

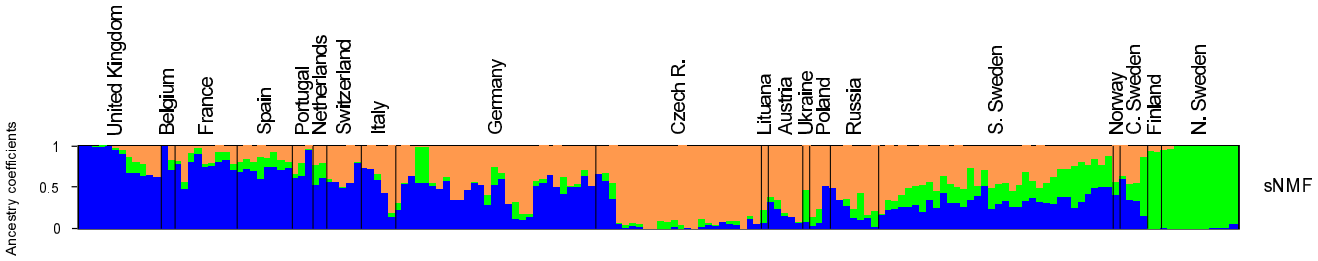

FIGURE 2.11: Représentation des coefficients de métissage pour 3 populations ancestrales pour la plante *Arapidopsis thaliana* (entropie croisée = 0.483).

2.1.3.4 Analyse de la structure de population chez *Arabidopsis thaliana*

Notre approche s'étant abstraite de l'hypothèse d'Hardy-Weinberg, cela nous permet d'estimer les coefficients de métissage pour des espèces consanguines. Nous avons pu appliquer le logiciel `sNMF` à l'espèce de plante *Arabidopsis thaliana* en Europe (Atwell et al., 2010). Le diagramme en barres de la matrice des coefficients de métissage, Q, montre une variation clinale le long de l'axe Est-Ouest de l'Europe séparant deux groupes, tandis que les individus scandinaves ont été placés dans un groupe séparé (Figure 2.11). Ces résultats sont proches d'estimations précédentes obtenues avec des données de séquences d'ADN (François et al., 2008).

2.1.3.5 Analyse de données du projet 1000 genomes

Afin d'illustrer l'utilité de notre approche, nous avons analysé les données issues de 1092 individus génotypés à plus de 2 millions de SNPs par le projet du consortium 1000 Genomes. Avec 5 populations ancestrales, le logiciel `sNMF` a identifié des groupes correspondant aux régions géographiques principales du monde (Figure 2.12). En particulier, un pourcentage considérable de métissage européen a été détecté chez les populations Afro-Américaines, chez les américains d'origine mexicaine, Portoricaines, et les Colombiennes par le logiciel `sNMF`.

2.1.3.6 Résumé

Nous avons présenté une nouvelle méthode fondée sur une factorisation de matrice non-négative pour estimer les proportions de métissage d'individus à partir de données génétiques. Cette approche s'appuie sur l'unification des approches de type "analyse en composantes principales"et de type "structure". La méthode proposée est 10 à 30 fois plus rapide que la méthode servant d'état-de-l'art sur les jeux de données que nous avons analysés. De plus,

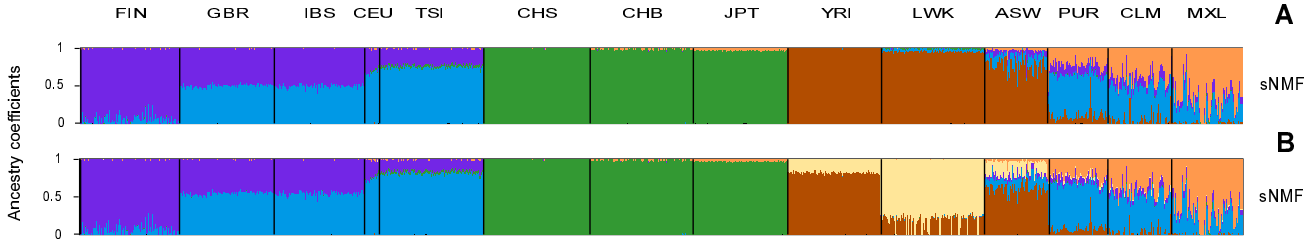

FIGURE 2.12: Représentation des coefficients de métissage pour le jeu du 1000 genomes pour 5 populations ancestrales (entropie croisée = 0.5010) et 6 populations ancestrales (entropie croisée = 0.5011). (FIN, Finnish ; GBR, British ; IBS, Spanish ; CEU, CEPH Utah residents ; TSI, Tuscan ; CHS, Southern Han Chinese ; CHB, Han Chinese ; JPT, Japanese ; YRI, Yoruba ; LWK, Luhya ; ASW, African-American ; PUR, Puerto Rican ; CLM, Colombian ; MXL, Mexican-American).

nous avons proposé un critère prédictif pour déterminer le nombre de populations ancestrales. Notre modélisation permet d'analyser des données issues d'espèces consanguines sans restriction particulière. Nous avons illustré notre approche par l'étude de plusieurs jeux de données humaines et d'un jeu de données de la plante *Arabidopsis thaliana* en Europe. Le détail de ce travail est présenté au chapitre 4.

2.2 Modèles Mixtes à Facteurs Latents (LFMM)

La détection de l'adaptation locale à partir de données génomiques représente une problématique importante pour les études écologiques. L'une des approches les plus utilisées pour réaliser une étude d'association écologique consiste à déterminer les SNPs corrélés avec un indicateur environnemental. Pour cela, de nombreuses méthodes de corrélation ont été développées. Nous avons fait l'état de l'art des principales méthodes au chapitre 1. Nous avons conclu qu'il est nécessaire de développer de nouvelles approches afin de réaliser des études d'association écologique permettant de détecter de l'adaptation locale chez une espèce, tout en corrigeant les biais dus aux facteurs de confusion impossibles à observer.

Dans cette partie, nous présentons une famille de modèles appelés *modèles mixtes à facteurs latents* (LFMM). Ce sont des modèles de régression utilisant la factorisation de matrice afin de modéliser les facteurs de confusion. Nous commençons par présenter le modèle, puis une comparaison avec différents logiciels servant d'état-de-l'art et enfin différentes analyses de données.

2.2.1 Modèle LFMM

Afin d'évaluer les associations entre les génotypes et d indicateurs environnementaux, $d \geq 1$, tout en corrigeant pour les facteurs de confusion, nous proposons de modéliser la matrice des génotypes centrée, $G^{(c)}$, comme une variable de réponse dans le modèle de régression bayésienne suivant

$$G_{i\ell}^{(c)} = U_i V_\ell^T + X_i B_\ell^T + \epsilon_{i\ell},$$

où B_ℓ est un vecteur des coefficients de regression (de dimension d), et X_i est un vecteur d'indicateurs environnementaux de dimension d. Nous supposons que B_ℓ est de loi a priori $N(0, D_B)$ avec D_B une matrice de covariance diagonale. Les termes U_i et V_ℓ sont des vecteurs de dimension K modélisant les facteurs de confusion. Le terme U_i est de loi a priori $N(0, \sigma_U^2 I_K)$ et le terme V_ℓ est de loi a priori $N(0, I_K)$. Le terme $\epsilon_{i\ell}$ est un terme de résidu de loi a priori $N(0, \sigma^2)$.

Nous utilisons un algorithme d'échantillonnage de Gibbs de la loi a posteriori des coefficients de régression, B, pour estimer la moyenne, l'écart type et le z-score associé à chaque coefficient.

Le modèle est appelé un *modèle mixte à facteurs latents* (LFMM). Par extension des modèles mixtes, le modèle LFMM corrèle les génotypes, G, avec les indicateurs environnementaux, X, tout en estimant, de manière conjointe, K facteurs latents U et leurs effets V. Ces K facteurs représentent les facteurs de confusion non observés. Le nombre de facteurs latents, K, est un paramètre important à déterminer. Nous proposons d'utiliser les estimations du nombre de populations ancestrales du logiciel `sNMF` ou le nombre de composantes significatives d'une ACP avec des tests de Tracy-Widom afin de trouver une valeur approchée du nombre de facteurs latents, K. Plusieurs valeurs de K sont à explorer systématiquement.

2.2.1.1 Modèles mixtes à facteurs latents dans la littérature

On retrouve des modèle similaires au modèle LFMM dans plusieurs domaines de la littérature. Sammel and Ryan (1996) ont étudié l'estimation par maximum de vraisemblance restreint sur des jeux de données de petites tailles. Un ensemble de modèles plus général que le modèle LFMM est les modèles d'équations structurelles (Sánchez et al., 2005). Des modèles similaires à LFMM, appelés modèles de régression à facteurs, ont aussi été considérés pour la construction de réseaux moléculaires à partir de données d'expression génique (West, 2003; Carvalho et al., 2008). Plus récemment, Woodard et al. (2013) ont décrit des modèles de régression à facteurs latents appliqués à l'épidémiologie.

Toutefois, notre approche se différencie de la littérature par plusieurs aspects. Tout d'abord, ces modèles ont souvent pour but d'interpréter les facteurs latents tandis que notre objectif est d'interpréter les effets fixes. De plus, le modèle LFMM a pour but d'effectuer un très grand nombre de tests pour un petit nombre d'individus, tandis que dans la littérature, les modèles ne sont pas utilisés pour effectuer des tests multiples, et ne prennent pas en compte le fléau de la grande dimension.

2.2.1.2 Simulations selon le modèle de LFMM

L'objectif principal de notre utilisation du modèle LFMM en génétique des populations est de déterminer les SNPs significativement associés à des indicateurs environnementaux. Pour cela, nous proposons un test fondé sur le calcul de $|z|$-scores. Nous avons vérifié la calibration du test de l'hypothèse nulle H_0 :"$B = 0$" dans le cadre de simulations issues du modèle génératif de LFMM.

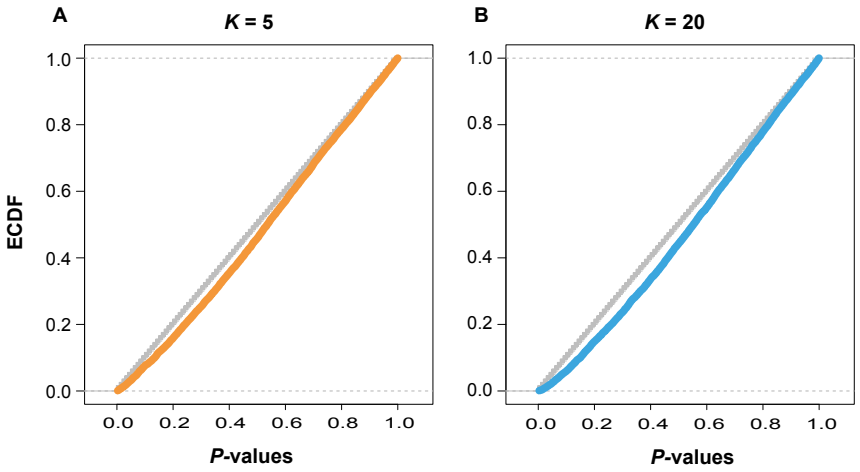

FIGURE 2.13: Répartition empirique des p-valeurs de LFMM en fonction d'une distribution uniforme, à partir de simulations selon le modèle de LFMM avec $B = 0$, A) $K = 5$ et B) $K = 20$.

Dans la figure 2.13, nous avons simulé des données selon le modèle de LFMM avec $B = 0$ pour différents nombres de facteurs latents ($K = 5$ et $K = 20$). Nous avons tracé la fonction de répartition empirique (ecdf) des p-valeurs en fonction de la distribution uniforme (Figure 2.13). Les répartitions empiriques des p-valeurs sont proches de la distribution uniforme. Les p-valeurs du test sont correctement calibrées dans les simulations.

2.2.2 Comparaisons avec l'état de l'art

2.2.2.1 Approches concurrentes

Nous proposons de comparer notre nouvelle approche avec un ensemble de méthodes faisant partie de l'état de l'art des méthodes d'association écologique. Parmi ces méthodes, on retrouve la régression linéaire (LM), la régression linéaire généralisée (GLM, Joost et al. 2007), le test de Mantel partiel (PMT, Smouse et al. 1986, Legendre and Legendre 2012) et les méthodes implantées dans les logiciels `GEMMA` et `BAYENV` (Zhou and Stephens, 2012; Coop et al., 2010).

Nous introduisons aussi le modèle PCRM (pour Principal Component regression model) de la manière suivante

$$G_{i\ell}^{(c)} = X_i B_\ell^T + \tilde{U}_i V_\ell^T + \epsilon_{i\ell},$$

K	LM	PCRM	LFMM
2	0.20	0.21	0.15
20	1.27	1.42	0.08
100	6.13	12.41	0.20

TABLE 2.2: Erreur des moindres carrés d'estimation des effets environnementaux

.

où $\tilde{U}_i$ représente les K premières composantes principales de la matrice G. La comparasion avec le modèle PCRM permet de savoir si l'apport du modèle LFMM provient de l'approximation de faible rang de la matrice des facteurs de confusion par la matrice $\tilde{U}V^T$ ou du fait que l'on estime conjointement B, U et V dans un cadre bayésien.

2.2.2.2 Simulations selon le modèle génératif de LFMM

Nous comparons la qualité de l'estimation du paramètre B pour la régression linéaire, pour les modèles PCRM et LFMM. L'objectif est de déterminer si le modèle LFMM apporte une meilleure estimation de B dans le cadre de simulations génératives.

Nous avons simulé des données selon le modèle génératif de LFMM pour $n = 100$ individus et $L = 1000$ locus avec $K = 2$, $K = 20$ et $K = 100$ facteurs latents. Nous avons calculé l'erreur moyenne des moindres carrés entre la valeur simulée de B et l'estimation des coefficients de régression de chaque modèle (Table 2.2). On observe une erreur d'estimation similaire pour les modèles LM, PCRM et LFMM avec $K = 2$ facteurs latents. Lorsque l'on augmente le nombre de facteurs latents à $K = 20$ puis $K = 100$, l'erreur reste du même ordre pour le modèle LFMM tandis que l'erreur croît d'un facteur 10 puis 100 pour les modèles LM et PCRM. Cela montre l'intérêt de la modélisation des facteurs latents lors de simulations avec le modèle génératif de LFMM.

2.2.2.3 Simulations de modèles de coalescence

Nous avons comparé le modèle LFMM aux modèles de l'état de l'art pour la détection d'associations écologiques dans le cadre des simulations utilisées en génétique des populations. Nous avons utilisé le logiciel `ms` (Hudson, 2002) pour simuler 200 individus et 1000 locus dans un modèle en îles. Ce modèle est composé de 40 îles, selon un habitat linéaire, modélisant l'isolement par la distance. Cela permet de mimer un jeu de données génétiques, sans sélection, comportant de la structure neutre entre les individus.

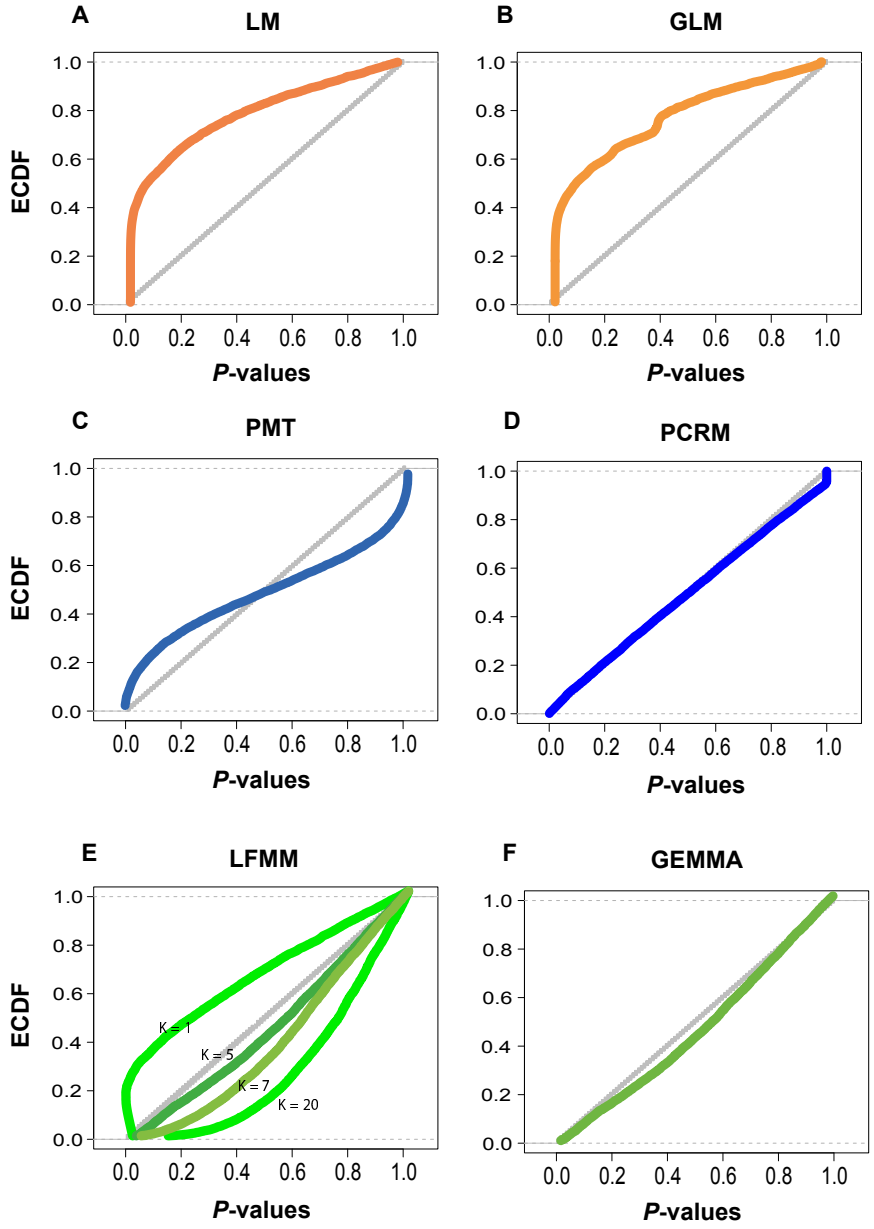

FIGURE 2.14: Répartition empirique des p-valeurs en fonction d'une distribution uniforme pour A) LM, B) GLM, C) PMT, D) PCRM, E) LFMM pour $K = 1, 5, 7, 20$, et F) GEMMA.

Nous nous sommes intéressés à la distribution des p-valeurs pour chacun des modèles (Figure 2.14). On remarque que pour les modèles LM et GLM, les p-valeurs sont sous-estimées. On a un grand nombre de faux positifs. Pour le test de Mantel partiel (PMT), la distribution des p-valeurs ne suit pas une distribution uniforme. Pour le modèle PCRM et le modèle du logiciel `GEMMA`, la distribution des p-valeurs est uniforme. Pour le modèle LFMM, on obtient une distribution proche de la distribution uniforme si l'on choisit $K = 5$ facteurs latents. On obtient un test libéral lorsque l'on choisit $K = 1$ facteur

FN (FP)	LM	GLM	PCRM	GEMMA	PMT	LFMM
type I error :						
$-\log_{10}\alpha = 3$	0% (33%)	0% (24%)	100 % (3%)	100 % (2%)	99% (6.8%)	4% (5%)
$-\log_{10}\alpha = 4$	0% (27%)	0% (19%)	100 % (0%)	100 % (0%)	100% (3.4%)	14% (3%)

TABLE 2.3: Pourcentage de Faux Négatifs (Faux Positifs) pour la régression linéaire (LM), la régression en composantes principales (PCRM), le modèle mixte de logiciel GEMMA, le test de Mantel partiel (PMT) et LFMM.

latent et un test conservatif lorsque l'on choisit $K = 7$ ou $K = 20$ facteurs latents. La distribution des p-valeurs pour le modèle LFMM dépend de la correction que l'on applique et plus précisément du nombre de facteurs latents que l'on utilise pour corriger les facteurs de confusion.

Il est important de noter que le choix du nombre de facteurs latents est difficile. Toutefois, on peut utiliser les méthodes que l'on a citées pour l'analyse de la structure de population (pourcentage de variance, test de Tracy-Widom, critère prédictif) pour s'aider dans ce choix. D'autres critères sont en cours d'évaluation, s'appuyant sur la statistique d'inflation génomique de Devlin and Roeder (1999).

Ensuite, nous avons évalué les différents algorithmes sur leur capacité à détecter des vraies associations. Pour cela, nous avons repris les simulations précédentes comportant de la structure neutre entre individus et nous y avons ajouté 50 locus binaires sous sélection. Ces locus ont été simulés avec une fréquence d'allèle, $f(x)$, variable en fonction de la coordonnée géographique de la population, x, selon la fonction suivante :

$$f(x) = \frac{1}{1 + e^{\theta(x-20)}}, \quad \theta > 0,$$

où $\theta = 0.1 - 0.2$ est la pente du gradient géographique de sélection (Haldane, 1948).

La Table 2.3 montre le pourcentage de Faux Négatifs (Faux Positifs) pour chacun des tests pour un seuil de significativité $\alpha = 10^{-3}$ et $\alpha = 10^{-4}$. On remarque que les modèles LM et GLM font un grand nombre de faux positifs mais sont capables de détecter l'ensemble des associations. Ces tests sont trop libéraux. À l'opposé, les tests des modèles du logiciel GEMMA, de PCRM et de PMT sont trop conservatifs. En effet, ces tests ne créent pas de faux positifs, mais ils détectent très peu d'associations réelles. Enfin, le modèle LFMM propose un compromis avec un faible nombre de faux positifs et un faible nombre de faux négatifs.

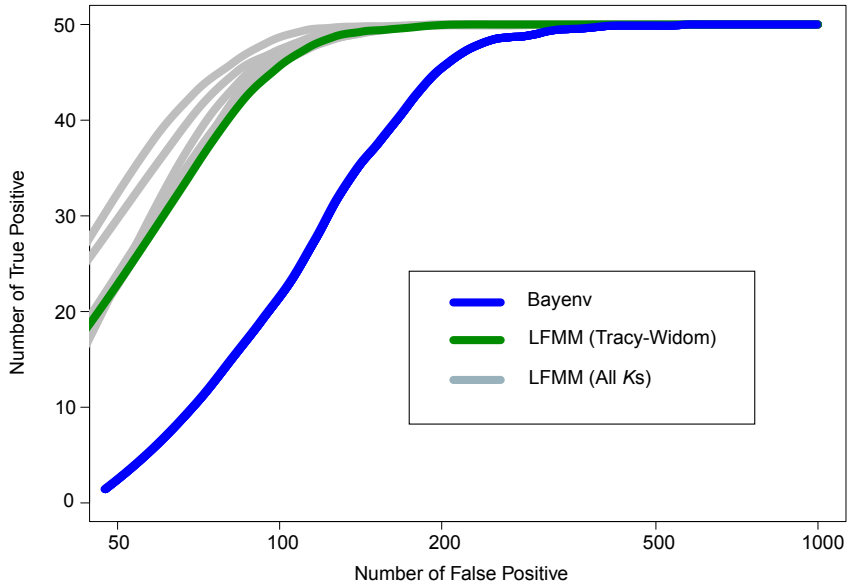

FIGURE 2.15: Courbe ROC pour le logiciel `LFMM` (vert, Tracy-Widom, gris sinon) et le logiciel `BAYENV` (bleu).

Nous avons ensuite comparé le modèle LFMM avec le modèle du logiciel `BAYENV` pour les mêmes simulations. Comme le logiciel `BAYENV` retourne des facteurs de Bayes et le logiciel `LFMM` des p-valeurs, il est difficile de les comparer directement. Nous comparons les 2 approches grâce à une courbe ROC pour laquelle nous avons ordonné les SNPs significativement associés à l'indicateur environnemental (Figure 2.15). Le logiciel `LFMM` détecte un plus grand nombre de vraies associations pour le même nombre de fausses associations en comparaison du logiciel `BAYENV`, quel que soit le nombre de facteurs considéré ($K = 1, 3, 5, 7, 10, 20$).

Par ailleurs, une étude de comparaison des modèles pour détecter les associations écologiques a été réalisée par De Villemereuil et al. (2014). Les modèles étudiés par De Villemereuil et al. (2014) sont la régression linéaire, le logiciel `BAYENV`, le logiciel `LFMM` et le logiciel `BAYESCAN` (Foll and Gaggiotti, 2008). D'après cette étude, `LFMM` propose le meilleur compromis entre puissance et taux de fausses découvertes dans les scénarios étudiés.

2.2.3 Application aux pins

Nous avons ensuite appliqué l'algorithme aux donnés génomiques collectées pour l'espèce de pin Loblolly (*Pinus taeda*, Eckert et al. 2010). Cette espèce de pin est distribuée dans tout le sud-est des États-Unis, allant de zones arides des plaines à des zones humides à

Table 3. Loblolly Pines.

Annotation	Gene Ontology	$-\text{Log}_{10}$(*P* Value)
Thylakoid lumenal 19 kDa chloroplast	Oxygen-evolving complex; Photosystem II	9.87
Pentatricopeptide repeat protein	Oxidative stress; salt stress	8.44
Conserved hypothetical protein	Ubiquitin-specific protease	8.28
Chalcone synthase	Flavonoid biosynthesis; wound response; oxidative stress	7.80
Heat shock	Temperature stress	7.67
Dirigent protein pdir18	Disease response	6.56
Heat shock transcription factor hsf5	Regulation of transcription; response to stress	6.15
Zinc finger	Transcription; DNA binding; zinc ion binding	5.84
Probable *n*-acetyltransferase hookless 1	Auxin signaling; photomorphogenesis; ethylene response	5.78
Calcium-binding pollen allergen	Polcalcin; calcium ion binding	4.61
Geranylgeranyl diphosphate synthase	Cholesterol biosynthesis; isoprenoid biosynthesis	4.59
Hypothetical protein Osl_04393	Trehalose-6-phosphate phosphatase	4.59
Potassium proton antiporter	Potassium ion transport; solute:hydrogen antiporter	5.54
DNA mismatch repair	DNA repair; regulation of DNA recombination	5.44

Note.—Annotation and gene ontology for some interesting SNPs with z-scores with absolute value greater than 4 for the first two components of 60 climatic variables.

Table 2.4: Liste de SNPs associés avec des gradients climatiques par `LFMM` et leur annotation fonctionnelle pour l'espèce de pin Loblolly

.

l'est du pays. Ce jeu de données a précédemment été étudié avec le logiciel `BAYENV` (Eckert et al., 2010).

Le logiciel `LFMM` permet de mettre en avant 392, 113 et 30 SNPs ayant un $|z|$-score supérieur à 4, 5 et 6 respectivement. Parmi les 50 SNPs les plus significativement associés à l'environnement, 17 ont aussi été déterminés par le logiciel `BAYENV`. La Table 2.4 donne une liste de SNPs associés avec des gradients climatiques par le logiciel `LFMM` et leur annotation fonctionnelle.

2.2.4 Application au jeu de données HGDP

Nous avons analysé avec le logiciel `LFMM` le jeu de données du projet HGDP composé de 1043 individus répartis en 52 populations et plus de 660000 SNPs (Li et al., 2008). Nous utilisons pour indicateur environnemental un résumé de variables climatiques (Hijmans et al., 2005). Le logiciel `LFMM` a permis de déterminer un ensemble de 2,624 SNPs (0.4 %) avec un $|z|$-score supérieur à 5. Parmi ces SNPs, 28 SNPs ont été découverts dans des études d'association à des traits ou à des maladies communes (Hindorff et al., 2009), et plusieurs SNPs ont été précédemment découverts par Hancock et al. (2011). Par exemple, les SNPs rs12913832 et rs28777 ont un $|z|$-score supérieur à 6 et sont associés aux gènes OCA2 et SLC45A2 (Table 2.5). Parmi les SNPs significativement associés à des gradients climatiques, plusieurs exemples se trouvent dans des gènes associés à la maladie cœliaque (ICOSLG), la taille (LHX3-QSOX2 et IGF1) et l'activation ou la synthèse de vitamine

Table 4. Human Data.

Landscape-Trait Category	Ref. SNP ID	Nearby Gene	Disease or Trait Association	$-\text{Log}_{10}$ (*P* Value)
Pigmentation and tanning	rs32579	*PPARGC1*	Tanning	9.42
	rs12913832	*OCA2/HERC2*	Eye color, eye color traits, hair color, black vs. blond hair color, black vs. red hair color	9.15
	rs11234027	*DHCR7*	Vitamin D levels	7.78
	rs3129882	*HLA-DRA*	Parkinson's disease	6.97
	rs28777	*SLC45A2*	Black vs. blond hair color, black vs. red hair color	6.90
Immune and autoimmune	rs1250550	*ZMIZ1*	Crohn's disease and inflammatory bowel disease (early onset)	8.77
	rs2735839	*KLK3*	Prostate cancer	8.16
	rs9264942	*RPL3P2*	HIV-1 control	8.02
	rs2179367	Intergenic between *SUMO4* and *ZC3H12D*	Dupuytren's disease	7.57
	rs1551398	Intergenic between *TRIB1* and *LOC100130231*	Crohn's disease	7.45
	rs2289700	*CTSH*	Bipolar disorder	6.98
	rs4819388	*ICOSLG*	Celiac disease	6.67
	rs703842	*CYP27B1/METTL1*	Multiple sclerosis	6.59
	rs12593813	*MAP2K5*	Restless legs syndrome	6.40
	rs4664308	*PLA2R1*	Nephropathy (idiopathic membranous)	6.28
Metabolism	rs10908907	Intergenic *MUC7*	Alcoholism (heaviness of drinking)	8.91
	rs1566039	Intergenic between *PAPD7* and *MIR4278*	Sphingolipid levels	6.89
	rs7665090	*MANBA*	Primary biliary cirrhosis	6.48
Cardiovascular	rs869244	*ADRA2A*	Platelet aggregation	7.20
	rs12034383	*CR1*	Erythrocyte sedimentation rate	7.15
	rs3129882	*HLA-DRA*	Systemic sclerosis	6.97
	rs11897119	*MEIS1*	PR interval	6.71
Height	rs7678436	*NCAPG-LCORL*	Height	9.43
Other	rs12479254	*BOK*	Brain structure	9.43

NOTE.—HGDP SNPs with the highest |z|-scores among those associated with phenotypic traits in GWAS.

TABLE 2.5: Annotations et les ontologies associées à un ensembles de SNPs détectés par le logiciel `LFMM`.

D (NADSYN1 et DHCR7) (Table 2.5). De plus, nous avons effectué une analyse d'enrichissement en ontologie sur les gènes contenant des SNPs avec un $|z|$-score supérieur à 5. Nous avons déterminé un enrichissement significatif dans des ontologies de gènes associées avec six processus biologiques liés à l'adhésion cellulaire, au déplacement cellulaire et au développement neuronal et organique.

2.2.5 Résumé

Nous avons proposé une nouvelle approche pour détecter l'adaptation locale au sein de génomes grâce au criblage génomique. Cette approche est fondée sur un modèle de régression modélisant les facteurs de confusion par des facteurs latents (LFMM). Nous avons montré que ce modèle est utile dans le cadre de simulations génératives et de simulations de génétique des populations. Ce modèle donne aussides résultats cohérents

avec les études précédentes lorsqu'il est appliqué à des données réelles. De plus, le logiciel `LFMM` a été utilisé, de manière indépendante, dans plusieurs études d'association écologique (Zueva et al., 2014; Kort et al., 2014; Stucki et al., 2014). Toutefois, il est important de rappeler que les méthodes statistiques que nous avons développées permettent seulement de proposer des mutations candidates. Des validations biologiques supplémentaires sont ensuite nécessaires. Une présentation plus détaillée du travail sur LFMM est disponible au chapitre 5.

2.2.6 Perspectives

Les questions en rapport avec les études d'associations écologiques sont nombreuses. Il existe donc de plusieurs pistes d'améliorations des méthodes proposées.

Une première piste pourrait être de prendre en compte la spécificité des données génomiques. En effet, la forte densité le long du génome des données considérées ne permet plus de négliger le déséquilibre de liaison, c'est-à-dire le lien entre les mutations proches sur le génome. De plus, il semble qu'il soit plus probable que les effets à rechercher soient faibles et polygéniques que forts et monogéniques (Zhou et al., 2013).

Une seconde piste pourrait être de chercher à mieux caractériser la part de chaque effet dans la variation génétique. Par exemple, on pourrait chercher à connaitre quelle part de la variation génétique est due aux pressions de l'environnement, quelle part est due à la structure des populations ou quelle part est due à l'isolement par la distance.

Une dernière piste pourrait être d'étendre le domaine d'application de modèle LFMM. LFMM est un modèle statistique général. Il pourrait être intéressant d'applique LFMM à des données phénotypiques ou morphologiques.

Chapitre 3

Correction des effets d'autocorrélation spatiale sur les cartes de composantes principales en génétique des populations

Dans de nombreuses espèces, la variation génétique spatiale montre des patrons d'isolement par la distance. Caractérisés par des fréquences d'allèles localement corrélées, ces motifs sont connus pour créer des formes périodiques dans les cartes géographiques de composantes principales. Ces formes influencent les interprétations des analyses en composantes principales (ACP). Dans cette étude, nous avons introduit des modèles regroupant un modèle d'ACP probabiliste et des modèles de krigeage. Ces modèles permettent d'estimer la structure génétique de population à partir de données génétiques tout en corrigeant pour les effets générés par des phénomènes d'autocorrélation spatiale. Ces algorithmes sont fondés sur une décomposition en valeurs singulières et une approximation de faible rang des données génotypiques. Comme leur complexité est proche de celle de l'ACP, ces algorithmes sont capables de passer à l'échelle avec la dimension des données. Pour illustrer l'utilité de ces nouveaux modèles, nous avons simulé des patrons d'isolement par la distance et de la variation géographique à grande échelle en utilisant des modèles de coalescence. Nos méthodes sont capables d'enlever les formes de fer à cheval que l'on observe habituellement dans des cartes de composantes principales et donc simplifient les interprétations de la variation génétique spatiale. Nous montrons l'utilité de notre approche par l'analyse d'un jeu de données de polymorphismes nucléotidiques issus du projet de diversité génomique humaine (HGDP) et nous comparons notre approche avec d'autres approches récemment développées.

Correcting principal component maps for effects of spatial autocorrelation in population genetic data

Eric Frichot, Sean D Schoville, Guillaume Bouchard and Olivier François.

Frontiers in Genetics. 2012 ; 3 :254.

Abstract

In many species, spatial genetic variation displays patterns of "isolation-by-distance". Characterized by locally correlated allele frequencies, these patterns are known to create periodic shapes in geographic maps of principal components which confound signatures of specific migration events and influence interpretations of principal component analyses (PCA). In this study, we introduced models combining probabilistic PCA and kriging models to infer population genetic structure from genetic data while correcting for effects generated by spatial autocorrelation. The corresponding algorithms are based on singular value decomposition and low rank approximation of the genotypic data. As their complexity is close to that of PCA, these algorithms scale with the dimensions of the data. To illustrate the utility of these new models, we simulated isolation-by-distance patterns and broad-scale geographic variation using spatial coalescent models. Our methods remove the horseshoe patterns usually observed in PC maps and simplify interpretations of spatial genetic variation. We demonstrate our approach by analyzing single nucleotide polymorphism data from the Human Genome Diversity Panel, and provide comparisons with other recently introduced methods.

Keywords

principal component analysis, isolation-by-distance, spatial autocorrelation, spatial factor analysis

3.1 Introduction

The concept of "isolation-by-distance" (IBD) was introduced by S. Wright to describe the accumulation of local genetic differences under spatially restricted dispersal (Wright, 1943). In species that are continuously distributed in geographic space and disperse over short distances, the theory predicts that genetic differentiation will increase with geographic distance (Malécot, 1948; Kimura and Weiss, 1964). IBD can be described by spatial autocorrelation, a measure of the degree of dependency among observations in a geographic space. Although studying IBD patterns could lead to useful estimates of gene dispersal (Rousset, 1997), spatial autocorrelation derived from IBD often presents a problem for population genetic analyses. More specifically, the presence of spatial autocorrelation patterns can increase the rate of false positive tests for hierarchical population structure or for the detection of loci under selection (Meirmans, 2012).

Recently, it has been acknowledged that distortions caused by spatial autocorrelation could also bias interpretations of population genetic structure as inferred from principal component analysis (PCA) or from Bayesian clustering methods (Novembre and Stephens, 2008; François et al., 2010). PCA is a method that searches for axes, called principal components, along which projected individuals show the highest variance. As a result, the first PCs are often used to explore the structure of variation in the sample. Characterized by locally correlated allele frequencies, IBD patterns create periodic shapes in PC maps that can confound signatures of migration events and influence interpretations of principal component analyses (Novembre and Stephens, 2008). In scenarios where covariance decays exponentially with geographic distance, PC plots are indeed expected to exhibit horseshoe effects, an artifact in which the second axis is curved relative to the first axis. These effects lead to counterintuitive representations of the data (Diaconis et al., 2008; Legendre and Gallagher, 2001).

Several methods have been proposed to correct for the effects of spatial autocorrelation in exploratory data analyses. In particular, those methods include spatial Principal Component Analysis (sPCA, Jombart et al. 2008; Borcard and Legendre 2002; Borcard et al. 2004; Dray et al. 2006), and sparse factor analysis (SFA, Engelhardt and Stephens 2010). Generally the methods share the objective of separating local and regional geographic scales in the data. In this study, we introduce a novel approach, based on latent factors models, that addresses the separation of geographic scales more directly than the two previous methods. The new method, spatial factor analysis (spFA), combines probabilistic PCA (Tipping and Bishop, 1999) and kriging models (Cressie, 1993) to infer population genetic structure from genetic data while correcting for errors introduced by spatial autocorrelation. While many approaches have been argued to improve interpretations of

the data, their outputs have not yet been compared to each other on the basis of spatial simulations. To compare methods, we generated patterns of IBD and broad-scale geographic variation using computer simulations of spatial coalescent models. We compared the outcomes of methods under population genetic models of isolation-by-distance, and we argued that the methods provided insights on distinct aspects of the data. We report that the new spFA method was able to remove the horseshoe effect observed in spatially structured data, whereas this was not the case in PCA, sPCA, and SFA analyses. We discuss the significance of this result in an assessment of single nucleotide polymorphism data from worldwide samples of the Human Genome Diversity Panel.

3.2 Materials and Methods

We considered single nucleotide polymorphism (SNP) data for n individuals genotyped at L loci. For these data, the genotypic matrix entries, $(G_{i\ell})$, record the number of derived alleles at locus ℓ for individual i. For autosomal data, $G_{i\ell}$ is thus equal to 0, 1, or 2, and corresponds to the genotype at locus ℓ. The data were centered by subtracting the mean value of each column of G and scaled by dividing by the standard deviation value of each column of G. In addition to the genotypic data, we assumed that geographical coordinates, (X_i), were recorded for each individual.

We evaluated the effects of IBD patterns on inference of population genetic structure using 4 statistical methods : Principal Component Analysis (PCA, Jolliffe 1986; Patterson et al. 2006), spatial PCA (sPCA, Jombart et al. 2008), Sparse Factor Analysis (SFA, Engelhardt and Stephens 2010), and a new method called *spatial Factor Analysis* (spFA).

3.2.1 Principal component analysis

PCA is a popular method that searches for a set of K orthogonal axes (the principal components), each of which is a linear combination of the original axes, such that projections of the original data display maximal variance onto the new axes (McVean, 2009). We computed the score matrix, U of dimension $n \times K$, and the loading matrix, V of dimensions $K \times L$, using the rank K singular value decomposition method implemented in the R function `prcomp` and in the computer program *SmartPCA* (Patterson et al., 2006).

3.2.2 Moran eigenvectors and spatial PCA

Moran eigenvectors maps were proposed as an alternative to trend surface analysis for incorporating spatial variation in population genetics models (Jombart et al., 2008; Dray et al., 2006). In Moran eigenvectors maps, there are positive and negative eigenvalues. Eigenvectors associated with positive eigenvalues have positive autocorrelation, and they describe global structures. Eigenvectors associated with negative eigenvalues describe local structures. Implemented in an algorithm called spatial PCA (sPCA), Moran's eigenvector maps (MEM) maximize Moran's spatial autocorrelation index, defined as follows

$$I(G) = \frac{\sum_{i,j} w_{ij}(g_i - \bar{g})(g_j - \bar{g})}{\sum_{i,j} w_{ij} \sum_i (g_i - \bar{g})^2}$$

with respect to a spatial weighting matrix, W, deduced from geographical distances and where g_i is the ith line of G (Dray et al., 2006). We implemented MEMs and sPCA using the R package `adegenet` using a Delaunay weighting matrix (Jombart et al., 2008).

3.2.3 Spatial factor analysis.

We introduce a new spatial factor analysis model (spFA) which incorporates spatial information in factor analysis in an explicit way. In spFA, inference is performed in a matrix factorization model similar to probabilistic PCA (Tipping and Bishop, 1999).

$$G_{i\ell} = U_i^T V_\ell + \epsilon_{i\ell}\,, \tag{3.1}$$

where $\epsilon_{i\ell}$ are statistically dependent Gaussian variables with mean zero and with covariance matrix Σ_θ. Similarly to Kriging approaches (Cressie, 1993), a radial basis covariance matrix was chosen to model spatial autocorrelation patterns generated by IBD (see also Durand et al. 2009). The covariance matrix Σ_θ was defined as follows. For all pairs of individuals, i and j, we have

$$\Sigma_\theta(i,j) = \exp(-d(X_i, X_j)/\theta), \quad \theta > 0\,, \tag{3.2}$$

where $d(X_i, X_j)$ represents the squared Euclidean or great-circle distance between sites with coordinate X_i and with coordinate X_j. To avoid collinearity issues, we assumed that the individual geographical coordinates were distinct from each other (ties were broken by adding small perturbations to the original spatial coordinates). The parameter θ is a scale parameter measured in units of average pairwise distance between geographic sites,

$\bar{d}$. In practice, spFA requires that an array of θ values (scale parameter) are explored, so θ was varied in the range $(0, 10\bar{d})$.

To solve the spFA model, we used a Cholesky decomposition, $C^TC = \Sigma_\theta^{-1}$, and we established an equivalence with the following matrix factorization model

$$\tilde{G}_{i\ell} = \tilde{U}_i^T \tilde{V}_\ell + \tilde{\epsilon}_{i\ell}\,, \tag{3.3}$$

where $\tilde{G} = CG$, $\tilde{U}^T = CU^T$, $\tilde{V} = V$, and where $\tilde{\epsilon_\ell}$ are statistically independent Gaussian vectors of mean zero and covariance matrix equal to identity. The matrix $\tilde{U}$ and $\tilde{V}$ were obtained by applying a singular value decomposition of rank K to the transformed data matrix, CG. Then, U and V were obtained by applying a singular value decomposition of rank K to $C^{-1}\tilde{U}^T\tilde{V}$. To avoid multiple solutions, the orthogonality condition $VV^T = I_K$, where I_K is the identity matrix in K dimensions, was imposed to V (Figure 3.1). The time needed to compute spFA is the same order as the time needed to compute K scores and loadings for a standard PCA (Patterson et al., 2006). For an example of implementation, see our R code (http ://membres-timc.imag.fr/Olivier.Francois/spfa.R).

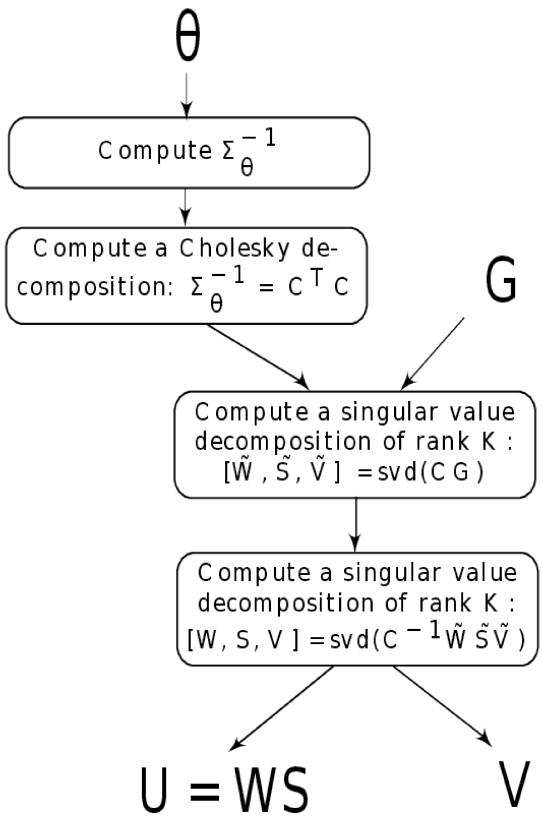

FIGURE 3.1: Algorithm for SpFA. For a genotypic matrix G with individual geographic coordinates (X_i), and for scale parameter $\theta > 0$, the spFA steps summarize as follows.

3.2.4 Sparse factor analysis

Sparse Factor Analysis (SFA) was introduced by Engelhardt and Stephens (2010) as an alternative to admixture-based models, and this method can recapitulate the results of PCA when population structure is influenced by IBD patterns. To give a description of SFA, we considered a regression model of the following form

$$G_{i\ell} = U_i^T V_\ell + \epsilon_{i\ell} \tag{3.4}$$

in which the residual errors are independent Gaussian random variables, $\epsilon_{i\ell} \sim N(0, 1/\psi_i)$, and where the prior distribution on the precision parameter, ψ_i, is a Gamma distribution. In the SFA model, an automatic relevance determination prior is considered for the score vectors, $U_{ik} \sim N(0, \sigma_{ik}^2)$, where some σ_{ik}^2 are constrained to be equal to zero. We implemented SFA using the code distributed in Engelhardt and Stephens (2010), and we used 1,000 iterations. Eigenvectors in spFA and in SFA are also referred to as *factors* or *axes*.

3.2.5 Simulated data

We generated simulated data for two diverging populations using coalescent models implemented in the computer program `ms` (Hudson, 2002). In these models, each population was simulated according to a linear stepping-stone model with 50 demes. To reproduce the simulation settings of Novembre and Stephens (2008), the effective migration rate between pairs of adjacent demes was set to the value $4Nm = 1$, where N is the population size and m is the migration rate. The divergence time τ between the two populations was varied within the range of values $\tau = (0, 100)$ measured in coalescent units. We sampled 100 individuals, one from each deme both side of a (fictive) geographic barrier. For each simulation, we evaluated Wilks' Λ, a statistic used in multivariate analysis of variance to test whether there are differences between the means of identified groups of individuals on the combination of genotypes (Mardia et al., 1979).

3.3 Results

3.3.1 Pure isolation-by-distance patterns

In a first series of experiments, we used simulations of one-dimensional stepping-stone models reproducing the patterns of IBD described in Novembre and Stephens (2008). In those simulated data, the divergence time between the two populations was thus set to $\tau = 0$, and the populations were connected by recurrent gene flow ($4Nm = 1$). As expected from theoretical results for PCA and for other ordination methods (Ahmed et al., 1974; Dray et al., 2006; Novembre and Stephens, 2008), the first PC maps displayed oscillating patterns. In addition, the frequency of oscillation increased as we examined axes of higher orders (Figure 3.2). When we used sPCA, the first three positive components were almost identical to those obtained with PCA (not reported).

Running spFA with $K = 3$ and with 3 distinct values of the scale parameter ($\theta/\bar{d} = 0.1, 0.2$ and 0.3) led to different interpretations of the genetic data (Figures 3.2B-D). Gradually varying θ allowed us to evaluate the scales at which the IBD effects were apparent, and also allowed us to remove those effects sequentially. For $\theta/\bar{d} = 0.1$, the maps corresponding to factor 1 and 2 displayed sinusoidal curves similar to PC1 and PC2, whereas the map for factor 3 was flat as expected if the effect of IBD is removed (Figure 3.2B). For $\theta/\bar{d} = 0.2$, the map corresponding to factor 1 remained similar to PC1, but the maps for factor 2 and factor 3 were flat (Figure 3.2C). For $\theta/\bar{d} = 0.3$, the effects of isolation-by-distance were corrected in all axes (Figure 3.2D).

When we ran SFA with $K = 3$ factors, the resulting maps also emphasized aspects of the data different from the ones described by PC maps and spatial factor maps (Figure 3.3). Maps for SFA are interpreted in terms of clusters, similar to those obtained in non-spatial Bayesian assignment programs like structure (Pritchard et al., 2000a). Clusters created by clustering programs under IBD models are often reported as being undesirable (François and Durand, 2010; Meirmans, 2012).

3.3.2 Two diverging populations with IBD patterns

In a second series of experiments, we used simulations of a two-population model, where each population consisted of a linear network of 50 demes. In these experiments, the two populations were separated by a geographic barrier to gene flow.

First the divergence time was set to $\tau = 10$ coalescent units. Using PCA, the first 2 components displayed oscillating patterns, similar to those obtained with $\tau = 0$ (pure

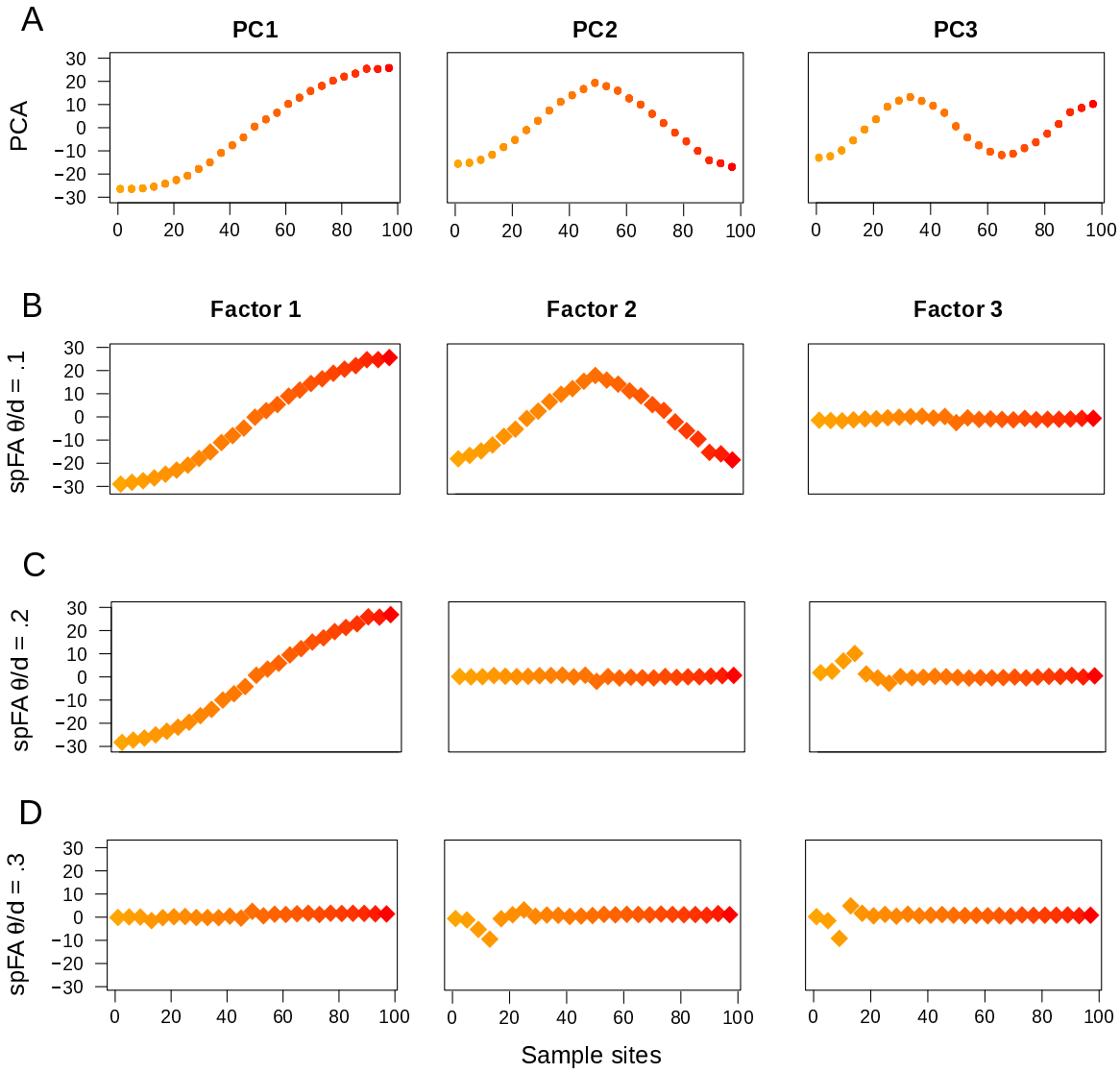

FIGURE 3.2: PC and spFA factor maps for data simulated under an IBD model. (A) PC maps, (B) spFA factor maps for $\theta/\hat{d} = 0.1$, (C) spFA factor maps for $\theta/\hat{d} = 0.2$, (D) spFA factor maps for $\theta/\hat{d} = 0.3$.

IBD simulations ; Figure 3.4A). The PC1-PC2 plot exhibited a clear horseshoe pattern. Differentiation between the two populations was visible in the PC1 map, where a discontinuity was observed at the center of the habitat. This discontinuity corresponded to the localization of the geographic barrier. Results for the positive eigenvectors of sPCA strongly resembled those obtained for the first PCs (Figure 3.4B).

Turning to spFA, we argued for a particular choice of $\theta/\bar{d}$ based on Wilks' Λ statistic, a standard measure of separation of groups in discriminant analysis, and computed this statistic for $\theta/\bar{d}$ ranging between 0.01 and 10. As spatial factor analysis provided different interpretations of the data depending on the scale at which the data were analyzed, the choice of θ was crucial to the method. Figure 3.5 reports the value of Wilks' Λ as a

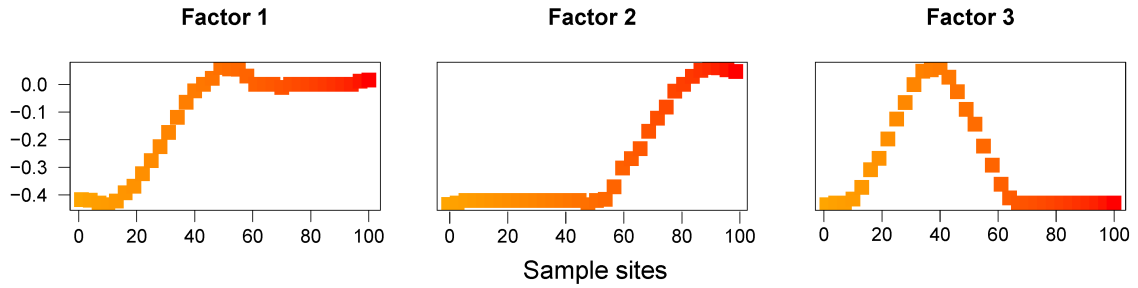

FIGURE 3.3: SFA for data simulated under an IBD model. Plots of the first three Factor maps for SFA.

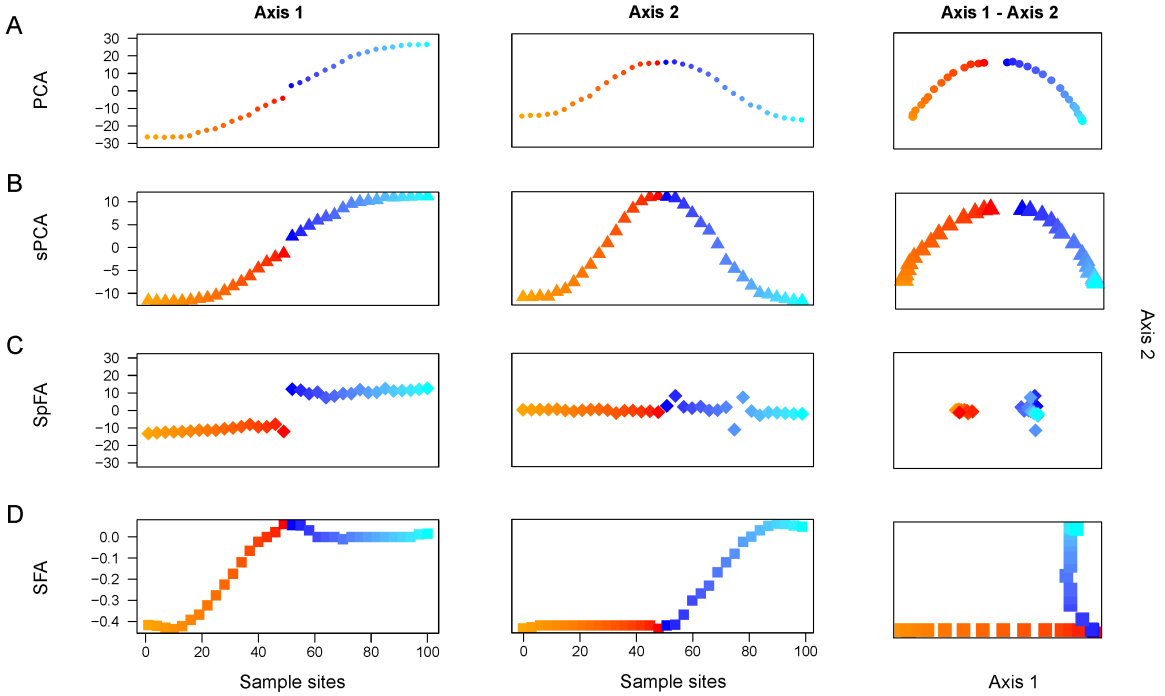

FIGURE 3.4: Two discrete populations under equilibrium IBD. Plots of the first 2 maps for (A) PCA, (B) sPCA, (C) spFA, (D) SFA.

function of the logarithm of $\theta/\bar{d}$. Values $\theta/\bar{d}$ minimizing Wilks' statistic and providing the best assignment of our data into clusters were about 0.32 (Figure 3.5). When spFA was applied with $K = 2$, the first factor map grouped demes at the left and the right of the geographic barrier in two main clusters, while simultaneously correcting for IBD patterns within the two clusters (Figure 3.4C). The spFA Axis1-Axis2 plot removed the horseshoe effect observed in PCA and sPCA plots. The resulting figure emphasized a discontinuous population structure consisting of two differentiated genetic clusters. Running SFA with $K = 2$ also led to a description of the data in two genetic clusters, located both sides of the geographic barrier, but the method failed to describe the two clusters as discontinuous

entities (Figure 3.4D).

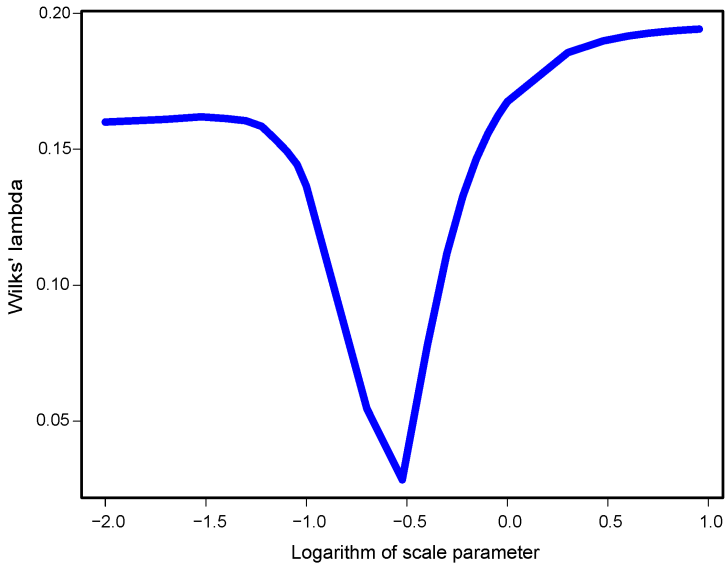

FIGURE 3.5: Wilks' Λ statistic as a function of the scale parameter $\theta/\bar{d}$ in spFA.

Based on PC and factor plots, we next computed Wilks' Λ statistic for all methods, and for divergence times τ ranging between 0 and 100 (Figure 3.6). Lower values of Λ generally indicated better discrimination of the 2 divergent populations in PC or factor plots. For all methods, the Λ statistic decreased as the divergence time between the 2 populations increased (McVean, 2009). In our spatially explicit framework, SFA (green curve) detected the existence of diverging populations earlier than PCA (red curve) and than sPCA (not shown, similar to PCA). SpFA was the most sensitive method, and provided an earlier detection of divergent clusters than SFA and PCA (blue curve).

3.3.3 Human data analysis

Next we applied PCA, sPCA, spFA, and SFA to a worldwide sample of genomic DNA from 418 individuals in 27 Asian populations, from the Harvard Human Genome Diversity Project - Centre Etude Polymorphism Humain (Harvard HGDP-CEPH) (ftp ://ftp.cephb.fr/hgdp_v3/). In those data, each marker has been ascertained in samples of Mongolian ancestry (referenced population HGDP01224). We selected all samples from Central and East-Asia with the exception of Xibe, which originated in northeastern China, but migrated to northwestern China only recently (Powell et al., 2007) (Figure 3.7A). The data set used a panel of 10,664 SNPs3 (see Patterson et al. 2012).

In our analysis, samples from Central Asia, west to the Tibetan plateau, were represented with red/orange colors, whereas populations from East-Asia were represented with blue

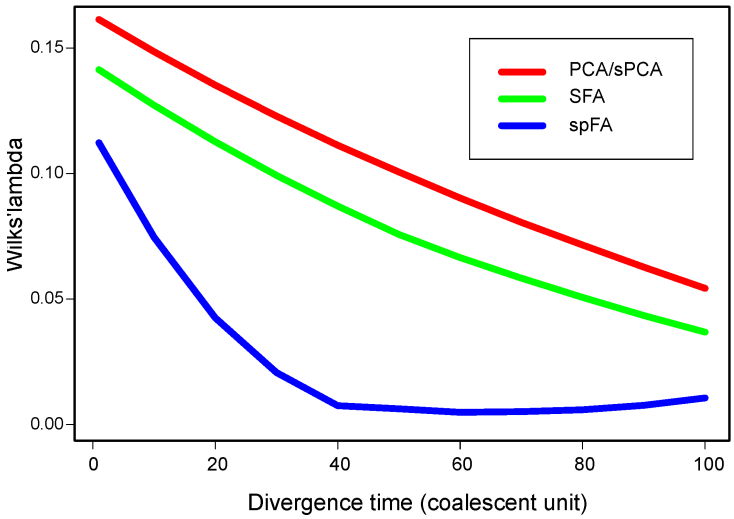

FIGURE 3.6: Wilks' Λ statistic as a function of the divergence time, τ, ranging between 1 and 100.

colors (Figure 3.7A). For those samples, the PC plot exhibited a horseshoe pattern, which was a signature of the presence of IBD patterns in the data (Figure 3.7B). PCA led to a continuum of samples without observable genetic discontinuities. Running spFA with $K = 2$ and setting $\theta/\bar{d} = 10^{-2}$ on the basis of Wilks' statistic analysis, spFA corrected for the effects of IBD in axes 1 and 2 (Figure 3.7C). The spFA method provided evidence of a major discontinuity separating two clusters, one in Central Asia and one in East-Asia. In addition, Uyghur and Hazara population samples aligned with the two main clusters and were placed in an intermediate position, suggesting genetic admixture from ancestral Central Asian and East-Asian gene pools. Essentially the same patterns emerged when spFA was applied with $K = 3$ at the same scale (Figures 3.8C-D).

Using SFA with $K = 2$, factors 1 and 2 confirmed the main discontinuity, in a representation of clusters closer to Bayesian clustering methods than to PCA (Figure 3.7D). Uyghur and Hazara population samples were also placed between the main clusters. When we used SFA with $K = 3$, we obtained shapes without natural interpretations (Figures 3.8A-B). SFA detected additional discontinuities whereas the other methods suggested that continuous genetic variation in geographic space was predominant.

3.4 Discussion

Principal component analysis and related methods used to describe genomic variation among large population samples are known to produce results that can be distorted by

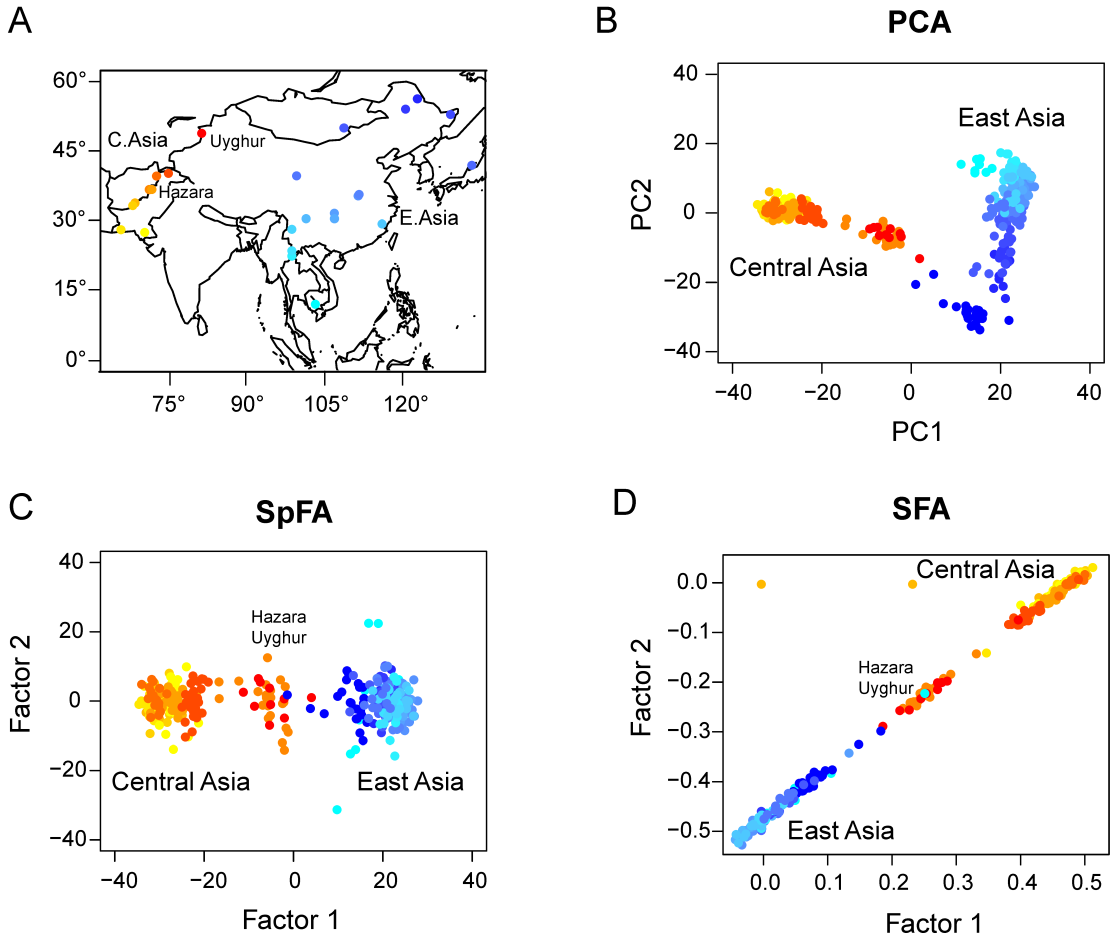

FIGURE 3.7: (A) Map of Asia with geographic locations of HGDP populations. PC and factor plots for (B) PCA, (C) spFA, (D) SFA.

IBD, and that may thus be difficult to interpret. The horseshoe effect is one of the distortions observed in PC plots that arises when covariance between allele frequencies decays exponentially with geographic distance. In this case, there is an established mathematical correspondence between the eigenvectors of the covariance matrix and the columns of a discrete cosine-transform (Ahmed et al., 1974; Diaconis et al., 2008). In this study, we used this correspondence to propose a new approach based on spatial models for the covariance structure of residual errors in factor analysis. In spFA, IBD effects were modeled through the introduction of a covariance matrix that accounts for the geographic distance between individuals explicitly.

We compared spFA to PCA and to two recent methods that also attempt to correct for IBD effects : spatial Principal Component Analysis (sPCA, Jombart et al. 2008) and sparse factor analysis (SFA, Engelhardt and Stephens 2010). When we applied PCA to

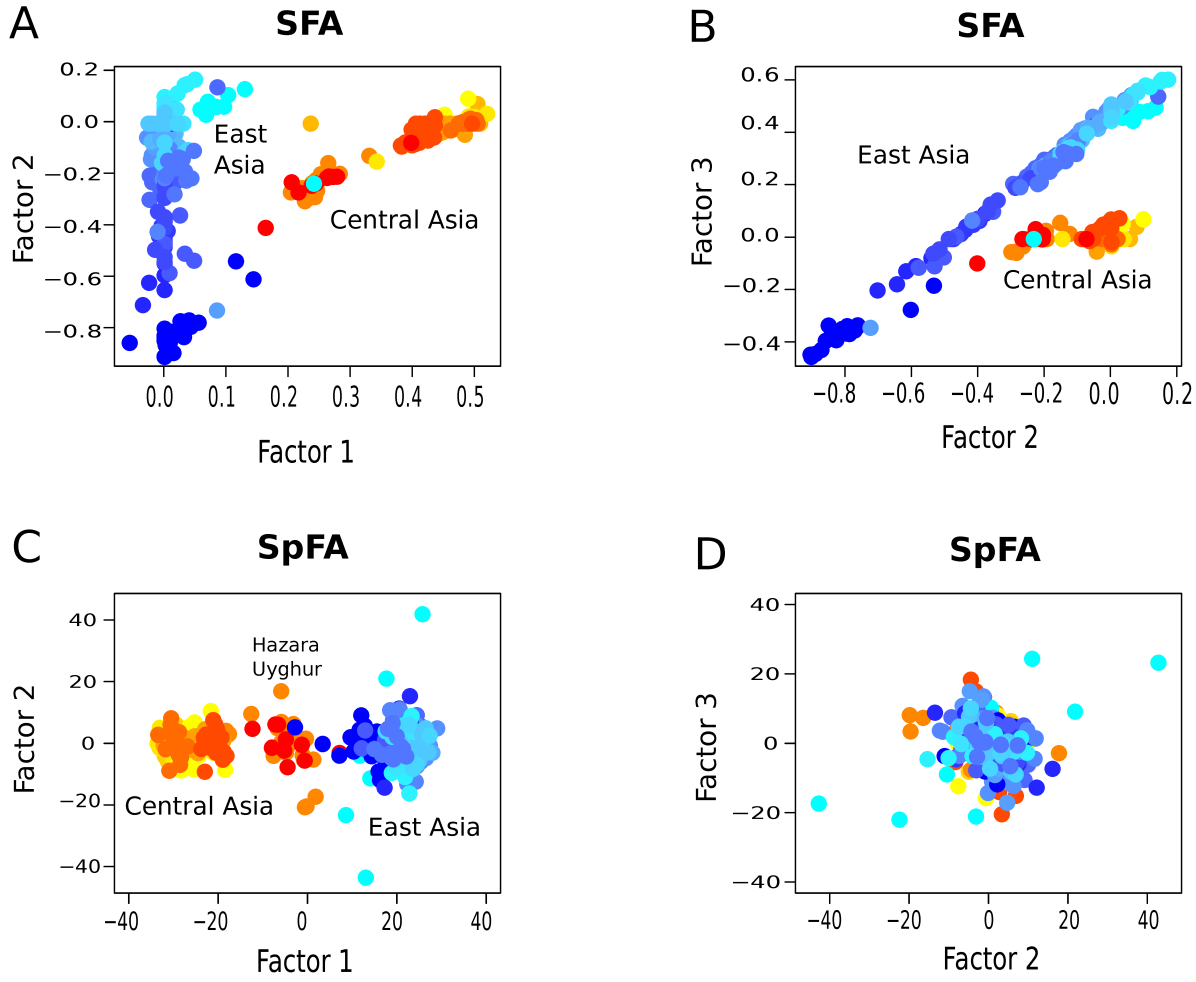

FIGURE 3.8: Factor plots for (A,B) SFA and (C,D) spFA with $K = 3$ clusters.

simulated data from spatial coalescent models, PC maps displayed sinusoidal curves as observed in previous studies (Novembre and Stephens, 2008). We observed that sPCA, which includes several distance matrices within Moran eigenvector maps of genetic data, produced results similar to those of PCA, and did not correct for IBD effects. When we applied SFA to spatial coalescent simulations, the algorithm clustered individuals in several small groups depending on the number of latent factors used in the method. SFA factor maps actually displayed outcomes closer to discrete clusters than to continuous variation. After adjusting for the spatial scale in the covariance model, spFA was able to remove the oscillating shapes observed in the first PCs sequentially.

When PCA was applied to spatially explicit simulations of two diverging populations, PC maps failed to firmly identify genetic discontinuities between populations. Despite a relatively long period of isolation in simulations, the populations were not strongly separated in PC maps due to the horseshoe effect. Compared to PCA and sPCA, the

spFA method had increased power to identify genetic discontinuities where they were masked by spurious autocorrelation effects. When we applied SFA, we found that, up to normalization of outputs, the results were similar to those generated by clustering algorithms like `structure`. For simulations of two diverging populations, SFA detected a main separation between two differentiated populations, but this approach did not correct for IBD effects within the main genetic clusters. Similarly to `structure`, the results of SFA were influenced by the presence of IBD patterns in the samples. We found that spFA alleviated this issue, and that it produced results more robust to the choice of the number of factors than SFA.

The methods used in this study provided quite distinct descriptions of the data when they were applied to human population samples from Central and East-Asia, and they underlined several aspects of the data. With PCA, a typical horseshoe pattern was observed, but no obvious genetic discontinuities were observed. In contrast, SFA provided evidence for two main clusters which were also confirmed by spFA. When we used SFA with $K = 3$, we obtained shapes without natural interpretations (Figure 3.8). SFA detected additional discontinuities whereas the other methods suggested that continuous genetic variation in geographic space was predominant. We observed that SFA behaves like clustering algorithms and did not correct for spurious clusters created by IBD patterns. This issue makes the SFA results difficult to interpret in terms of admixture and ancestral populations. The spFA method corrected for the horseshoe pattern observed in PC plots by removing autocorrelation effects from the second and third axes. The method suggested that Asian population structure is strongly influenced by IBD patterns. In the spFA plot, Hazara of Pakistan and Uygur of northwestern China grouped together, and were placed between Pakistani and East-Asian populations (Rosenberg et al., 2002). These results either support the presence of admixed genomes in Hazara and Uygur populations, or favor the hypothesis of a central Asian migration route of modern humans in East-Asia (Zhang et al., 2007). The public availability of data sets other than the HGDP will enable us to further assess the utility of the method for analyzing human genetic data.

A potential limitation of the spFA approach is it's sensitivity to the choice of the scale parameter, θ. The θ parameter actually determines the scale of the spatial effects that could be removed by spFA. Note that spFA is essentially performing a standard principal component analysis when it is applied with small values of the scale parameter. In this study, we recommended exploring a grid of θ values so that IBD effects could be removed at distinct scales sequentially. The choice of the number of factors, K, in spFA is also tied to the particular value of θ implemented in the model. One way to determine K is by using Tracy-Widom tests on the matrix of genotypes, $\tilde{G}$ (Patterson et al., 2006). Gradually increasing the value of θ enabled a fine grain analysis of genetic discontinuities in human

data, and allowed us to study IBD patterns within genetic clusters. The computational complexity of spFA increases linearly as a function of the number of markers. Since it is equivalent to the computation of a low rank approximation of the genotypic matrix (lower than a standard PCA, a few seconds on standard computer systems), applying spFA at multiple scales was not overly time-consuming.

3.4.1 Conclusion

This study provided a comparison of existing methods that attempt to correct for IBD effects in population genetic analyses, and showed that each of studied approaches provided different insights on the data. Under equilibrium IBD, PCA was confounded by continuous variation and the main genetic discontinuities may be missed or misinterpreted. For the same data, SFA over-estimated the number of clusters in the genetic data, creating spurious clusters from continuous patterns. In the presence of IBD patterns, spatial factor analysis provided clearer interpretations of the data than PCA and SFA. In a spatially explicit framework, we found that spFA identified genetic discontinuities more efficiently than did PCA or SFA when these discontinuities are blurred by noise from IBD patterns in the genetic data.

3.5 Acknowledgments

We thank Nicolas Duforet-Frebourg for his help with the software `ms`. This work was supported by a grant from la Région Rhône-Alpes to Eric Frichot and Olivier François, and by an NSF grant to Sean Schoville (OISE-0965038). Olivier François acknowledges support from Grenoble INP.

Chapitre 4

Estimation rapide et efficace des coefficients de métissage individuel

L'estimation de coefficients de métissage individuel est importante pour la génétique des populations et les études d'associations. Cette estimation est habituellement faite en utilisant des algorithmes coûteux en calcul s'appuyant sur la maximisation de la vraisemblance. Avec l'apparition de jeux de données génomiques de grande taille, la recherche de versions rapides d'algorithmes de maximisation de la vraisemblance a généré une activité considérable. Cependant, réduire le coût de calcul de tels algorithmes représente toujours un défi majeur. Dans cet article, nous présentons une méthode rapide et efficace pour estimer les coefficients individuels de métissage fondée sur des algorithmes de factorisation de matrice non-négative. Nous avons implanté notre méthode dans le programme `sNMF` et nous l'avons appliquée à des jeux de données humaines et de plantes. Les performances du logiciel `sNMF` ont ensuite été comparées avec celles de l'algorithme de maximum de vraisemblance implanté dans le logiciel `ADMIXTURE`. Sans perte de précision, le logiciel `sNMF` calcule les estimations des coefficients de métissage 10 à 30 fois plus rapidement que le logiciel `ADMIXTURE`.

Fast and efficient estimation of individual ancestry coefficients

Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, Olivier François.

Genetics. 2014 ; 196 :973–983.

Abstract

Inference of individual ancestry coefficients, which is important for population genetic and association studies, is commonly performed using computer-intensive likelihood algorithms. With the availability of large population genomic data sets, fast versions of likelihood algorithms have attracted considerable attention. Reducing the computational burden of estimation algorithms remains, however, a major challenge. Here, we present a fast and efficient method for estimating individual ancestry coefficients based on sparse nonnegative matrix factorization algorithms. We implemented our method in the computer program `sNMF` and applied it to human and plant data sets. The performances of `sNMF` were then compared to the likelihood algorithm implemented in the computer program `ADMIXTURE`. Without loss of accuracy, `sNMF` computed estimates of ancestry coefficients with runtimes $\approx 10 - 30$ times shorter than those of `ADMIXTURE`.

Keywords

inference of population structure, ancestry coefficients, nonnegative matrix factorization algorithms

4.1 Introduction

Inference of population structure from multilocus genotype data is commonly performed using likelihood methods implemented in the computer programs `STRUCTURE`, `FRAPPE`, and `ADMIXTURE` (Pritchard et al., 2000a; Tang et al., 2005; Alexander et al., 2009). These programs compute probabilistic quantities called ancestry coefficients that represent the proportions of an individual genome that originate from multiple ancestral gene pools. Estimation of ancestry proportions is important in many respects, for example in delineating genetic clusters, drawing inference about the history of a species, screening genomes for signatures of natural selection, and performing statistical corrections in genome-wide association studies (Pritchard et al., 2000b; Marchini et al., 2004; Price et al., 2006; Frichot et al., 2013).

Individual ancestry coefficients can be estimated using either supervised or unsupervised statistical methods. Supervised estimation methods use predefined source populations as ancestral populations. Classical supervised estimation approaches were based on least-squares regression of allele frequencies in hybrid and source populations (Roberts and Hiorns, 1965; Cavalli-Sforza and Bodmer, 1971). Unsupervised approaches attempt to infer ancestral gene pools from the data, using likelihood methods. An undesired feature of likelihood methods is that they can be computer intensive, with typical runs lasting several hours or more. With the use of dense genomic data and increased sample sizes, reducing the time lag necessary to perform estimation is a major challenge of population genetic data analysis.

A fast approach to the estimation of ancestry coefficients is by using principal component analysis (PCA, Patterson et al. 2006). PCA is an exploratory method that describes high-dimensional data, using a small number of dimensions, and makes no assumptions about sampled and ancestral populations. Using PCA can lead to results surprisingly close to likelihood methods, and connections between methods have been intensively investigated during recent years (Patterson et al., 2006; Engelhardt and Stephens, 2010; Frichot et al., 2012; Lawson et al., 2012; Lawson and Falush, 2011). But a drawback of PCA is that interpretation in terms of ancestry is often difficult, as it can be confounded by demographic factors or irregular sampling designs (Novembre and Stephens, 2008; McVean, 2009; François and Durand, 2010).

In this study, we introduce computationally fast algorithms that lead to estimates of ancestry coefficients comparable to those obtained with `STRUCTURE` or `ADMIXTURE`. The algorithms were implemented in the computer program sNMF based on sparse nonnegative matrix factorization (NMF) and least-squares optimization (Lee and Seung, 1999;

Kim and Park, 2007, 2011). Like PCA, NMF algorithms are flexible approaches that are robust to departures from traditional population genetic model assumptions. In addition, NMF algorithms produce estimates of ancestry proportions with runtimes that are much shorter than those of `STRUCTURE` or `ADMIXTURE`. This study assesses the utility of NMF algorithms when analyzing population genetic data sets and compares the performances of the algorithms implemented in sNMF with those implemented in `ADMIXTURE` on the basis of human and plant data.

4.2 Materials and Methods

To provide statistical estimates of ancestry proportions using multilocus genotype data sets, we implemented sparse NMF least-squares optimization algorithms in the computer program `sNMF`.

4.2.1 Modeling ancestry coefficients

We considered allelic data for a sample of n multilocus genotypes at L loci representing single-nucleotide polymorphisms (SNPs). The data were stored into a genotypic matrix (X), where each entry records the number of derived alleles at locus ℓ for individual i. For autosomes in a diploid organism, the number of derived alleles at locus ℓ is then 0, 1, or 2. In our algorithm, we used 3 bits of information to encode each 0, 1, or 2 value as an indicator of a heterozygote or a homozygote locus. In other words, the value 0 was encoded as 100, 1 was encoded as 010, and 2 as 001. The use of a binary coding warrants that the entries sum up to L for each row of the transformed data matrix.

Admixture models generally suppose that the genetic data originate from the admixture of K ancestral populations, where K is unknown a priori. Given K populations, the probability that individual i carries j derived alleles at locus ℓ can be written as

$$p_{i\ell}(j) = \sum_{k=1}^{K} q_{ik} g_{k\ell}(j) \quad j = 0, 1, 2, \tag{4.1}$$

where q_{ik} is the fraction of individual i's genome that originates from the ancestral population k, and $g_{k\ell}(j)$ represents the homozygote $(j = 0, 2)$ or the heterozygote $(j = 1)$ frequency at locus ℓ in population k. Since it makes no assumption about Hardy–Weinberg equilibrium, the above framework is appropriate to deal with inbreeding and outbreeding

in ancestral populations. Using our binary coding, Equation 1 writes as

$$P = QG\,, \tag{4.2}$$

where $P = (p_{i\ell})$ is an $n \times 3L$ matrix, $Q = (q_{ik})$ is an $n \times K$ matrix, and $G = (g_{k\ell}(j))$ is a $K \times 3L$ matrix. The Q matrix records ancestry proportions for each individual in the sample. Although the focus of the above framework is on estimating ancestry estimates for each sampled individual, it can be easily modified to provide ancestry estimates based on allele frequencies in population samples.

4.2.2 Least-squares estimates of ancestry proportions

We approached the inference of ancestry coefficients by using least-squares (LS) optimization algorithms (Engelhardt and Stephens, 2010). Estimates of the Q and G matrices were obtained after minimizing the least-squares criterion

$$\mathrm{LS}(Q, G) = \|X - QG\|_{\mathrm{F}}^2\,, \tag{4.3}$$

where $\|M\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix M (Berry et al., 2007). Without constraints on Q and G, the solutions of the LS problem are given by the singular value decomposition of the matrix X, and the resulting matrices Q and G contain the scores and loadings of a PCA. To obtain ancestry coefficients, the matrices Q and G must have nonnegative entries such that

$$\sum_{k=1}^{K} q_{ik} = 1\,, \quad \sum_{j=0}^{2} g_{k\ell}(j) = 1. \tag{4.4}$$

With the constraints of Equation 4, estimating ancestry coefficients and genotypic frequencies is equivalent to performing NMF of the data matrix, X. NMF was previously applied to gene expression data (Kim and Park, 2007), and algorithms for NMF were surveyed and compared in Kim and Park (2011). In `sNMF`, estimates of Q and G were computed using the alternating nonnegativity-constrained least-squares (ANLS) algorithm with the active set (AS) method (Berry et al., 2007; Kim and Park, 2011). We modified the ANLS-AS algorithm as follows.

Our algorithm begins with the initialization of the Q entries with nonnegative values. Then it iterates the following cycles until convergence. The first step of an algorithm

cycle consists of computing a nonnegative matrix G that minimizes the quantity

$$\mathrm{LS}_1(G) = \left\| X - QG \right\|_F^2 , G \geq 0. \tag{4.5}$$

The G matrix was obtained by setting all negative entries to zero, after solving classical linear regression equations. The obtained solution was then normalized so that its entries satisfy Equation 4.

Given G, the second step of the cycle consists of computing a nonnegative matrix Q that minimizes the quantity

$$\mathrm{LS}_2(Q) = \left\| \begin{pmatrix} G^T \\ \sqrt{\alpha}\ e_{1\times K} \end{pmatrix} Q - \begin{pmatrix} X^T \\ 0_{1\times n} \end{pmatrix} \right\|_F^2 \tag{4.6}$$

where $e_{1\times K}$ is a row vector having all entries equal to 1, $0_{1\times n}$ is a vector of length n with all entries equal to 0, and α is a nonnegative *regularization parameter*. This minimization problem was solved using the block principal pivoting method proposed by Kim and Park (2011). The obtained solution, Q, was then normalized so that the row entries sum up to 1. Iterations were stopped based on a stationarity criterion derived from the Karush–Kuhn–Tucker conditions (Kim and Park, 2011) and when the relative difference between two successive values of the criterion was less than a tolerance threshold of $\epsilon = 10^{-4}$.

For $\alpha > 0$, the algorithm amounts to performing *sparse* NMF for the data matrix X. We tested values $\alpha > 0$ because they can reduce the variance of Q and G estimates for the smaller data sets, force irrelevant estimates to zero, and improve the numerical behavior the ANLS minimization algorithm. In addition, the programming structures used in `sNMF` optimized the time spent in memory access. Several algorithmic methods were also used to accelerate computation of matrix products. While we evaluated `sNMF` runtimes using a single computer processor unit (2.4 GHz, 64 bits, Intel Xeon), a multithreaded version of the `sNMF` program was also developed for multiprocessor systems.

4.2.3 Data sets

Ancestry inference and runtime analyses were performed on six worldwide samples of genomic DNA from 52 populations of the Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain (HGDP-CEPH). Five panels were extracted from the Harvard HGDP-CEPH database. These panels were given identification nos. HGDP00778,

Data set	Sample size	No. SNPs	Reference
HGDP00778	934	78,000	Patterson *et al.* (2012)
HGDP00542	934	48,500	—
HGDP00927	934	124,000	—
HGDP00998	934	2,600	—
HGDP01224	934	10,600	—
HGDP-CEPH	1,043	660,000	Li *et al.* (2008)
1000 Genomes	1,092	2,200,000	1000 Genomes Project Consortium (2012)
A. thaliana	168	216,000	Atwell *et al.* 2010)

TABLE 4.1: Data sets used in this study.

HGDP00542, HGDP00927, HGDP00998, and HGDP01224 and contained precisely ascertained genotypes of $n = 934$ individuals. The genotypes were specifically designed for population genetic analyses (Patterson et al., 2012). Each marker was ascertained in individuals of Han, Papuan, Yoruba, Karitiana, and Mongolian ancestry, and the data matrices included 78,253, 48,531, 124,115, 2635, and 10,664 SNPs, respectively (Patterson et al. 2012, Table 4.1). A sample of 1043 individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel was also analyzed. The genotypes were generated on Illumina 650K arrays (Li et al., 2008), and the SNP data were filtered to remove low-quality SNPs included in the original files. In addition, we used data from the 1000 Genomes Project. The 1000 Genomes Project data contain the genomes of 1092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing [phase 1 data 1000 Genomes Project Consortium. 2012]. The data matrix included 2.2 million polymorphic sites across the human genome (Table 4.1).

To examine the robustness of `sNMF` to departures from classical population genetic hypotheses, additional analyses were performed on a sample of n = 168 European accessions of the plant species *Arabidopsis thaliana*. *A. thaliana* is a widely distributed self-fertilizing plant known to harbor considerable genetic variation and complex patterns of population structure and relatedness (Atwell et al., 2010). We analyzed 216,130 SNPs spread across the genome of *A. thaliana* (Atwell et al. 2010, Table 4.1).

4.2.4 Comparisons with `ADMIXTURE`

The computer program `ADMIXTURE` (version 1.22) estimates ancestry coefficients based on the likelihood model implemented in `STRUCTURE`. In `ADMIXTURE`, the assumption of Hardy–Weinberg equilibrium in ancestral populations translates into a binomial model for allele counts at each locus. Considering unrelated individuals, the logarithm of the likelihood can thus be computed as

$$\mathcal{L}(Q,F) = \sum_i \sum_\ell \left(x_{i\ell} \log(\sum_k q_{ik} f_{k\ell}) + (1 - x_{i\ell}) \log(\sum_k q_{ik}(1 - f_{k\ell})) \right)$$

up to an additive constant that does not influence estimation algorithms. In this formula, $Q = (q_{ik})$ represents the matrix of ancestry coefficients for all individuals, and $F = (f_{k\ell})$ represents a matrix of allele frequencies for all loci. The F matrix can be converted to a G matrix comparable to the one computed by `sNMF`, using the binomial model, $g_{k\ell}(0) = (1-f_{k\ell})^2$, $g_{k\ell}(1) = 2f_{k\ell}(1-f_{k\ell})$, and $g_{k\ell}(2) = f_{k\ell}^2$. `ADMIXTURE` provides numerical estimates of Q and F that maximize the quantity $\mathcal{L}(Q, F)$. The local optimization algorithm relies on a block relaxation scheme, using sequential quadratic programming for block updates, coupled with a quasi-Newton acceleration of convergence.

A difficulty with optimization algorithms used by `ADMIXTURE` and `sNMF` is that the solutions produced can be dependent on the initial values used for Q, F, or G. To enable comparisons with estimates obtained with `ADMIXTURE`, the clusters output by runs of each program was permuted using `CLUMPP` (Jakobsson and Rosenberg, 2007). Differences in ancestry estimates obtained with `ADMIXTURE` (Q^{ADM}) and with `sNMF` (Q^{sNMF}) were assessed by two measures. The first measure was defined as the root mean-squared error (RMSE) between the matrices Q^{ADM} and Q^{sNMF} obtained from each program

$$\text{RMSE} = \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} (q_{ik}^{\text{ADM}} - q_{ik}^{\text{sNMF}})^2 \right)^{1/2}.$$

Although G matrices could be mainly considered as nuisance parameters for our estimation problem, a similar RMSE criterion was defined for comparing them. The second measure was defined as the squared Pearson correlation coefficient (R^2) between the matrices Q^{ADM} and Q^{sNMF}. When simulations with known Q matrices were analyzed, one of the two matrices was replaced by the true Q matrix used to generate the simulated data.

Runs of `ADMIXTURE` and `sNMF` were performed for values of the number of clusters set to $K = 2 - 10$, 15, and 20 for human data sets and set to $K = 2 - 7$ for *A. thaliana*. For `sNMF`, the values of the regularization parameter (α) ranged between 0 and 10,000, using a $\log_{10}$ scale (5 values). Each run was replicated five times for a total of 1410 experiments. Missing data imputation was initially performed after resampling missing genotypes from empirical frequencies at each locus. The missing values were updated using predictive probabilities after 20 sweeps of the algorithm (see below).

4.2.5 Cross-entropy criterion

We employed a cross-validation technique based on imputation of masked genotypes to evaluate the prediction error of ancestry estimation algorithms (Wold, 1978; Eastment and Krzanowski, 1982). The procedure partitioned the genotypic matrix entries into a

training set and a test set. To build the test set, 5% of all genotypes were randomly selected and tagged as missing values. The occurrence probabilities for the masked entries were computed using the program outputs obtained from training sets according to the formula

$$p_{i\ell}^{\text{pred}}(j) = \sum_{k=1}^{K} q_{ik} g_{k\ell}(j)\,, \quad j = 0, 1, 2\,. \tag{4.7}$$

`ADMIXTURE` predicts each masked value by $E[x_{i\ell}|Q^{\text{ADM}}, F^{\text{ADM}}] = 2\sum_k q_{ik}^{\text{ADM}} f_{k\ell}^{\text{ADM}}$ and the prediction error is estimated by averaging the squares of the deviance residuals for the binomial model (Alexander and Lange, 2011). Extending the approach employed by `ADMIXTURE` to our nonparametric approach, the predicted values were compared to the masked values, $x_{i\ell}$, by averaging the quantity defined as $-\log p_{i\ell}^{\text{pred}}(x_{i\ell})$ over all SNPs in the test set. In statistical terms, our criterion provides an estimate of the quantity

$$H(p^{\text{sample}}, p^{\text{pred}}) = -\frac{1}{nL} \sum_{i\ell} \sum_{j=0}^{2} p_{i\ell}^{\text{sample}}(j) \log p_{i\ell}^{\text{pred}}(j)\,, \quad j = 0, 1, 2. \tag{4.8}$$

This quantity corresponds to the sum of the Kullback–Leiber divergence between the sampled (p^{sample}) and predicted (p^{pred}) allelic distributions and the Shannon entropy of the sample distribution. It also corresponds to the cross-entropy between p^{sample} and p^{pred}. The number of ancestral gene pools (K) and the regularization parameter (α) were chosen to minimize the cross-entropy criterion. In general, smaller values of the criterion indicate better algorithm outputs and estimates. The standard error of the cross-entropy criterion is of order $1/\sqrt{n_L}$ where n_L is the number of masked genotypes. For data sets including 1000 individuals genotyped at $> 20,000$ SNPs, the third digit of the cross-entropy criterion can be significant.

4.2.6 Simulated data analysis

We adopted a simulation approach to compare RMSEs between the Q matrix computed by `ADMIXTURE` or by `sNMF` and a known matrix used to generate the simulated data. In addition, we assessed whether the correct value of K could be identified by `sNMF`, using the cross-entropy criterion.

In a first series of simulations, we used the 1000 Genomes Project data set to generate artificial data showing various levels of admixture. As ancestral populations, we chose the Han Chinese (CHB), British (GBR), and Yoruba (YRI) samples (1000 Genomes Project Consortium., 2012). We considered 50,000 SNPs in linkage equilibrium, exhibiting no missing genotypes. The allele frequencies observed in our three ancestral populations

were used as the true values for the F matrix. The genotypic matrix was constructed according to the binomial model used by `ADMIXTURE`. For 1000 individuals in each simulated data set, a Q matrix was simulated from a Dirichlet probability distribution, and several parameters were explored. Our experiments reproduced the parameters used for evaluating the accuracy of `ADMIXTURE` in a previous study (Alexander et al., 2009). Runs of `sNMF` were performed for values of the number of clusters set to $K = 2 - 5$ ($\alpha = 0$), and the choice of K was made on the basis of the cross-entropy criterion. For $K = 3$, the values of the regularization parameter (α) were varied between 0 and 100.

Additional data sets were created to mimic the population structure of European populations of *A. thaliana*, using 10,000 SNPs (168 individuals). To define ancestral frequencies, we used the western European populations, grouping samples from the United Kingdom, Belgium and France (23 individuals) ; central European populations, grouping samples from the Czech Republic (24 individuals) ; and northern European populations, grouping samples from Finland and Northern Sweden (13 individuals). `ADMIXTURE` and `sNMF` grouped these samples within three well-separated clusters exhibiting low levels of admixture with other plant populations. The empirical frequencies computed from the three populations were considered as the true frequencies for a generative model with $K = 3$ ancestral populations. From empirical frequencies, we computed genotypic frequencies, $f_{k\ell}$, using four distinct values of population inbreeding coefficient, $F_{\text{IS}} = 25 - 100\%$, that corresponded to moderate and strong levels of inbreeding. For 168 individuals, 10,000 genotypes were simulated using the sampling equation $p(x_{i\ell} = j) = \sum_k q_{ik} g_{k\ell}(j)$ where q_{ik} corresponds to the Q matrix computed from the full empirical data set (216,000 SNPs). In addition, simulated data sets were generated with or without missing data (0 or 20%). Fifty replicates were created for each value of the inbreeding coefficient and for each value of the ratio of missing data.

4.3 Results

We used the program `sNMF` to implement nonnegative matrix factorization algorithms and to compute least-squares estimates of ancestry coefficients for worldwide human population samples and for European populations of the plant species *A. thaliana*. As in the likelihood model implemented in the computer programs `STRUCTURE` and `ADMIXTURE`, `sNMF` supposes that the genetic data originate from the admixture of K parental populations, where K is unknown, and it returns estimates of ancestry proportions for each multilocus genotype in the sample (Pritchard et al., 2000a; Alexander et al., 2009). To

estimate ancestry coefficients, `sNMF` solves a constrained least-squares minimization problem, using an alternating algorithm based on a block principal pivoting method (Kim and Park, 2011) (see Materials and Methods, section 4.2).

4.3.1 Comparison of ancestry estimates for HGPD data sets

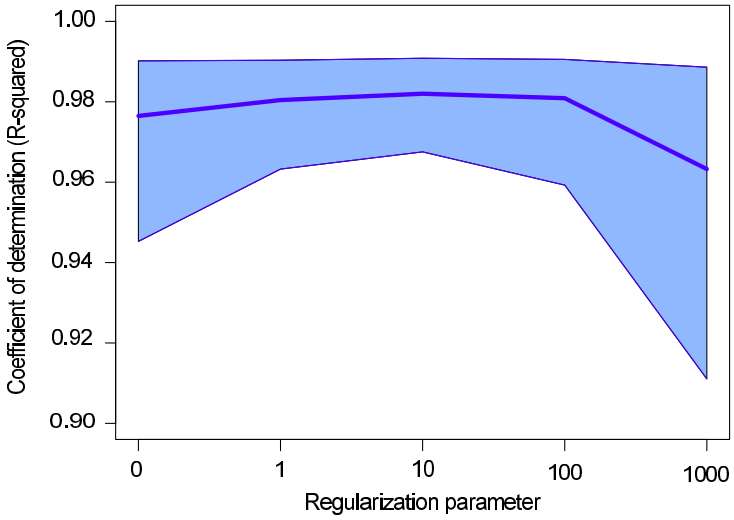

FIGURE 4.1: Correlation between `sNMF` and `ADMIXTURE` estimates. Squared correlation (coefficient of determination, R^2) between the ancestry coefficients estimated by each program. For each regularization parameter, the result corresponds to the maximum correlation over 5 runs, averaged over the number of clusters (ranging from 5 to 10) and over 6 HGDP data sets. The shaded area corresponds to a 95% confidence interval displayed for each value of the regularization parameter, α

.

First we evaluated the ability of `ADMIXTURE` estimates to be accurately reproduced by `sNMF` for five Harvard HGDP panels and for the HGDP-CEPH data set (Li et al., 2008; Patterson et al., 2012). For each run of `ADMIXTURE`, we computed a maximum squared correlation coefficient (R^2) and a minimum RMSE over runs of `sNMF` performed with the same number of clusters (K). For K ranging from 5 to 10, squared correlation coefficients remained > 0.96 across all runs (480 runs, Figure 4.1). Average values of the RMSE remained $< 5.5\%$ across all runs (Supporting Information, Supplementary Table 4.2). These results provided evidence that `sNMF` estimates closely reproduce those obtained with `ADMIXTURE` across the six HGDP data sets.

α	0	1	10	100	1000
RMSE	0.046 [0.035,0.064]	0.044 [0.035,0.057]	0.041 [0.035,0.052]	0.041 [0.031,0.061]	0.055 [0.033, 0.095]

TABLE 4.2: Root mean square error (RMSE) for 6 HGDP data sets as a function of the regularization parameter.

4.3.2 Run-time analysis

		$K = 5$		$K = 10$		$K = 20$	
Data set (no. SNPs)	Time unit	sNMF	ADMIXTURE	sNMF	ADMIXTURE	sNMF	ADMIXTURE
HGDP01224	min	0.68	4.4	0.8	11	1.7	48
(10,600)		[0.1, 1.6]	[3.4, 4.9]	[0.18, 1.7]	[9.9, 12]	[1.3, 1.8]	[41, 55]
HGDP00778	hr	0.087	0.61	0.12	1.5	0.25	6.2
(78,000)		[0.03, 0.15]	[0.55, 0.66]	[0.044, 0.12]	[1.3, 1.5]	[0.14, 0.34]	[3.8, 9.4]
HGDP-CEPH	hr	0.9	3.7	0.92	12	2.1	38
(660,000)		[0.33, 1.3]	[3, 4.3]	[0.38, 1.5]	[11, 12]	[1.3, 3.0]	[29, 45]
1000 Genomes Project	hr	2.8	(19)	4.6	(59)	—	—
(2,200,000)		[1.1, 4.7]	—	[1.5, 8.3]	—	—	—

Terms in brackets represent 95% confidence intervals. Terms in parentheses represent values obtained from a single program run.

TABLE 4.3: Run-time summary for `sNMF` and `ADMIXTURE`. Average values and their 95% confidence intervals.

Next we performed runtime analyses for `ADMIXTURE` and for `sNMF`, using the 1000 Genomes Project phase 1 data in addition to the previous HGDP data sets. The runtimes were averaged over distinct random seed values for each value of K. Runtimes increased with the number of SNPs in the data set and with the number of clusters in each algorithm (Figure 4.2, Table 4.3, Supplementary Figure 4.3). For data set HGDP01224 (10,600 SNPs), it took on average 0.8 min (1.7 min) for `sNMF` to compute ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the runtime was on average 11 min (48 min) for $K = 10$ ($K = 20$) clusters. For panel HGDP00778 (78,000 SNPs), it took on average 7.2 min (15 min) for `sNMF` to compute ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the average runtime was 1.5 hr (6.2 hr) for $K = 10$ ($K = 20$) clusters. For the CEPH-HGDP data sets (660,000 SNPs), it took on average 55 min (2.1 hr) for `sNMF` to compute ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the average runtime was 12 hr (38 hr) for $K = 10$ ($K = 20$) clusters. Runtimes increased in a quadratic fashion with K for `ADMIXTURE` whereas they increased linearly for `sNMF` (Figure 4.2). For the values of K used in our analyses, `sNMF` ran $5 - 30$ times faster than `ADMIXTURE` when these programs were applied to HGPD data sets. Regarding the 1000 Genomes Project phase 1 data set, the average runtimes of `sNMF` were ≈ 2.8 hr (4.6 hr) for $K = 5$ ($K = 10$) clusters. The `ADMIXTURE` runs led to similar estimates of Q, but a single run on the phase 1 data set took > 19 hr for $K = 5$ (59 hr for $K = 10$).

4.3.3 Prediction of masked genotypes

To decide which program options could provide the best estimates, we employed a cross-validation technique based on the imputation of masked genotypes (Wold, 1978; Alexander and Lange, 2011). The cross-validation method partitions the genotypic matrix entries into a training set and a test set that are used for estimation and validation sequentially.

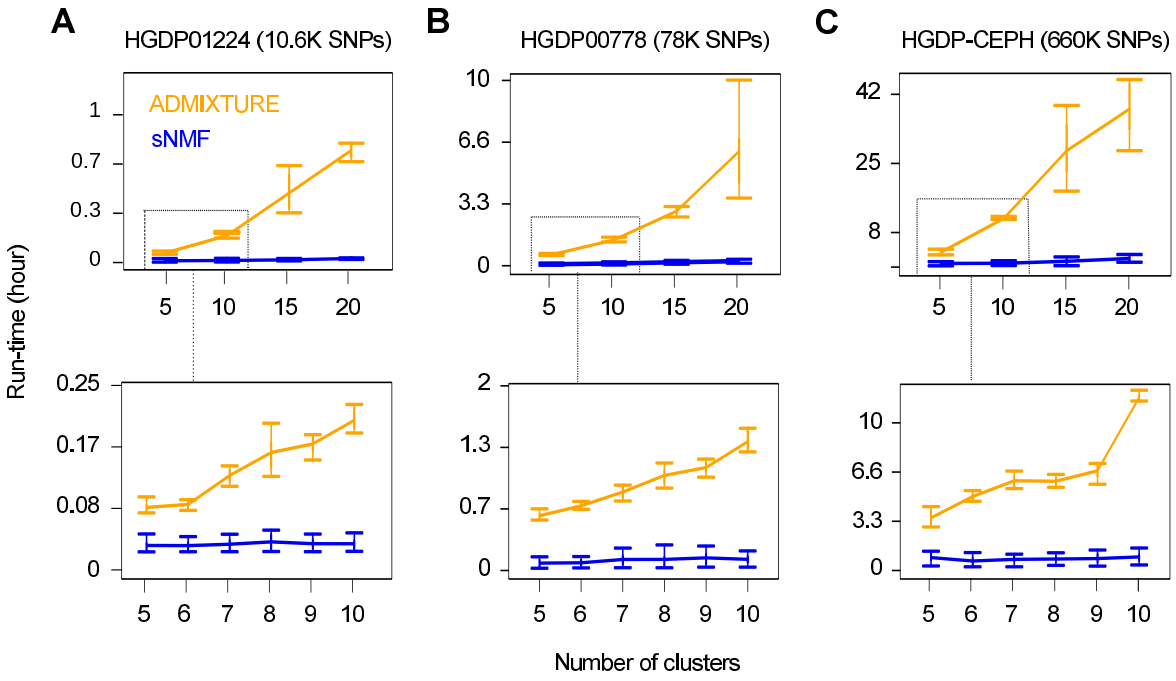

FIGURE 4.2: Runtimes for `sNMF` and `ADMIXTURE` runs. Averaged time elapsed before the stopping criterion of the `sNMF` (blue) and `ADMIXTURE` (orange) programs is met. Time is expressed in unit of hours. (A) Runtime analysis for Harvard HGDP panel 01224 (10,600 SNPs). (B) Runtime analysis for Harvard HGDP panel 00778 (78,000 SNPs). (C) Runtime analysis for the HGDP-CEPH data (660,000 SNPs).

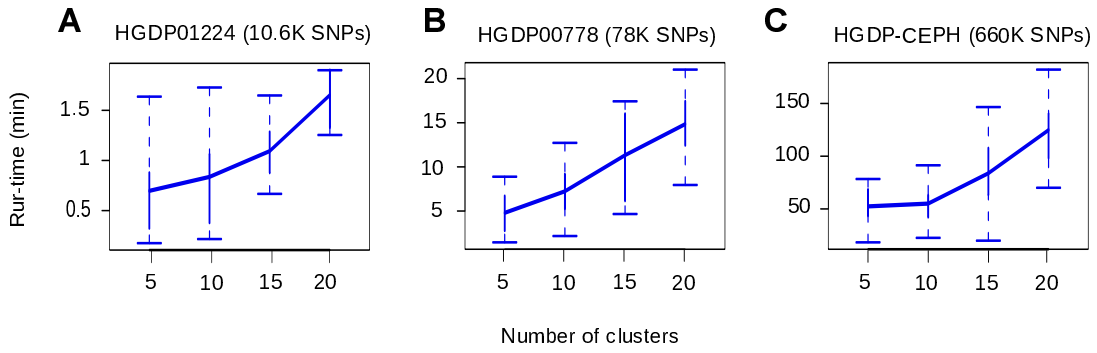

FIGURE 4.3: Runtimes for `sNMF`. Time is expressed in unit of minutes. A) Run-time analysis for Harvard HGDP panel 01224 (10.6K SNPs). B) Run-time analysis for Harvard HGDP panel 00778 (78K SNPs). C) Run-time analysis for the HGDP-CEPH data (660K SNPs)

To build test sets, 5% of the genotypic matrix entries were tagged as missing values. The masked entries were then predicted using estimates obtained from training sets. Predictions were assessed using a cross-entropy criterion that measured the capability of an algorithm to correctly impute masked genotypes (see Materials and Methods, section 4.2).

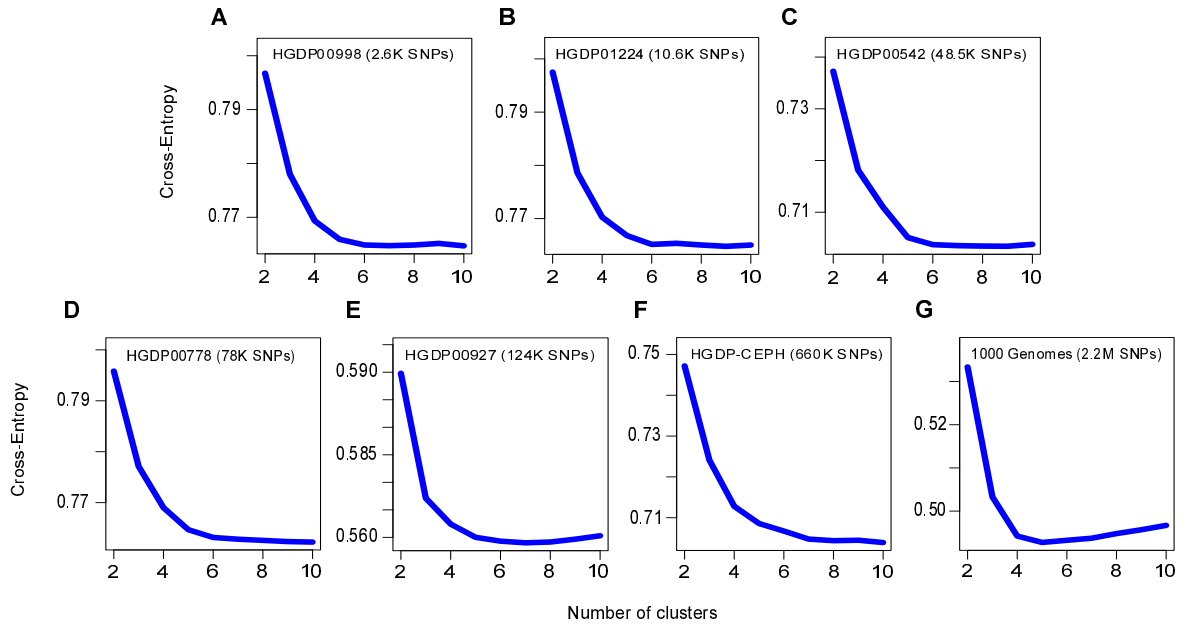

FIGURE 4.4: Values of the cross-entropy criterion for `sNMF` runs (human data sets). (A–G) Minimal values of the cross-entropy criterion over five runs of `sNMF` for (A–E) five Harvard HGDP panels, (F) HGDP-CEPH data, and (G) the 1000 Genomes Project data. The number of clusters ranged from 2 to 10.

Lower values of the cross-entropy criterion generally indicate better predictive capabilities of an algorithm.

Using the cross-entropy criterion, we performed an extensive analysis of `sNMF` program outputs to assess which values of the number of clusters (K) and the regularization parameter (α) could provide the best prediction of masked genotypes (Figure 4.4, Supplementary Figure 4.5). For HGDP data sets with moderate size (panels HGDP00998 and HGDP01224), values of $K \approx 7 - 8$ provided the best predictive results. For larger human data sets, cross-entropy values did not stabilize for $K \leq 10$, indicating that > 10 clusters were necessary to describe population structure. Choices of regularization parameter values > 1000 were generally discarded by the cross-entropy criterion. For panels of moderate size, the best ancestry estimates were obtained for values $\approx \alpha = 100$. The influence of the regularization parameter was substantial for the smallest data sets, but for the largest ones a wide range of values led to comparable imputation results (Supplementary Figure 4.5). Regardless of the value of the regularization parameter, $K = 5$ clusters led to the best results for the 1000 Genomes Project data set (Figure 4.4). This last result is in accordance with the criteria used for choosing populations included in the 1000 Genomes Project.

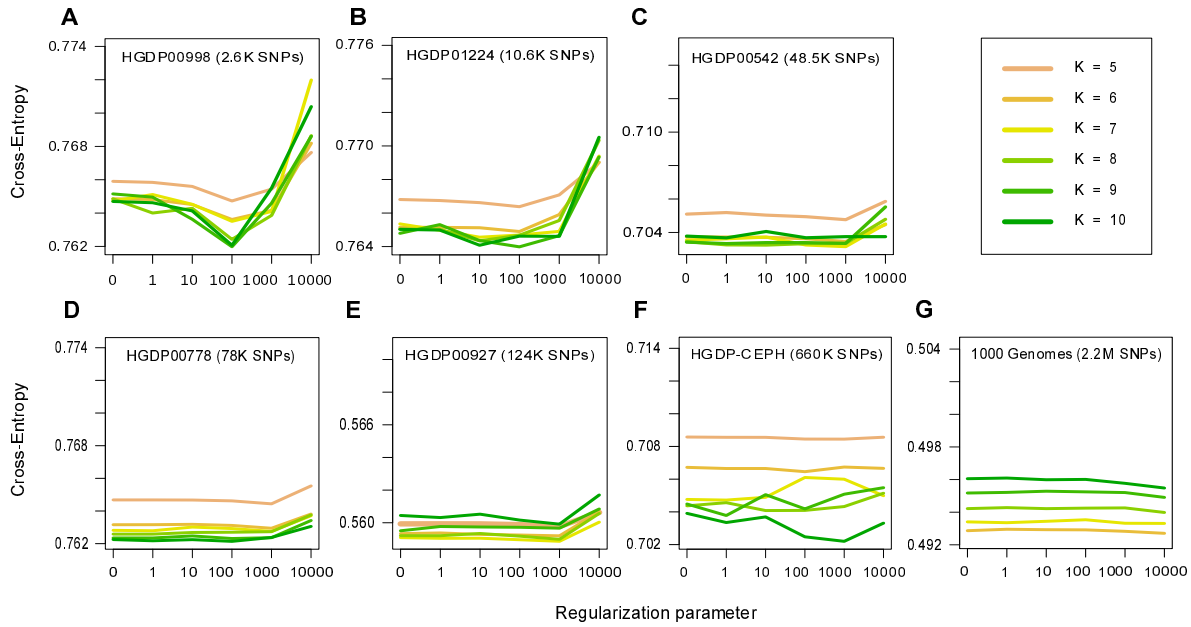

FIGURE 4.5: Values of the cross-entropy criterion for `sNMF` algorithms. Minimal values of the cross-entropy criterion over 5 runs of the `sNMF` program for A-E) 5 Harvard HGDP panel, F) the HGDP-CEPH data, and G) The 1000 Genomes Project data set. The number of clusters ranged from 5 to 10, and the values of the regularization parameter ranged from 0 to 10,000.

4.3.4 Ancestry estimates

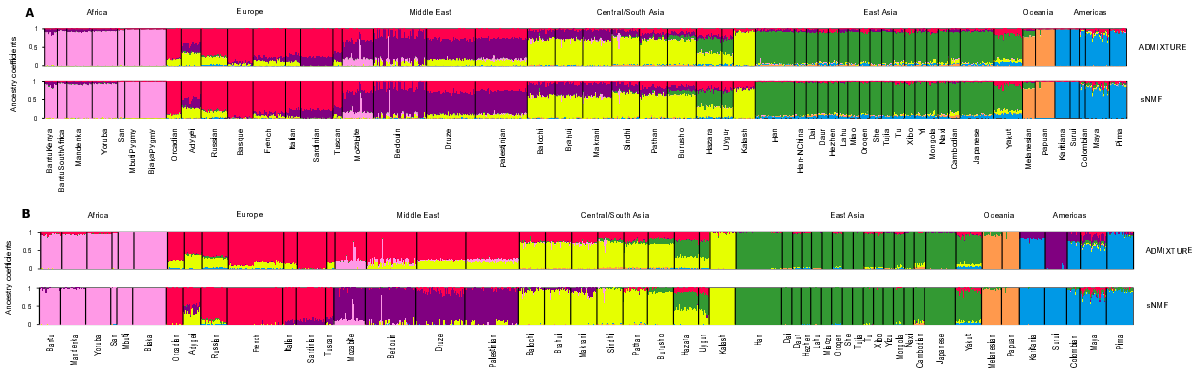

FIGURE 4.6: Graphical representation of ancestry estimates obtained for HGDP data sets ($K = 7$). (A) HGDP00778 panel (78,000 SNPs). Shown are estimated ancestry coefficients using `ADMIXTURE` (top, cross-entropy = 0.747) and `sNMF` (bottom, cross-entropy = 0.762 and $\alpha = 100$). (B) HGDP-CEPH data set (660,000 SNPs). Shown are estimated ancestry coefficients using `ADMIXTURE` (top, cross-entropy = 0.691) and `sNMF` (bottom, cross-entropy = 0.704 and $\alpha = 100$).

To compare ancestry estimates obtained from particular runs of `sNMF` and `ADMIXTURE`,

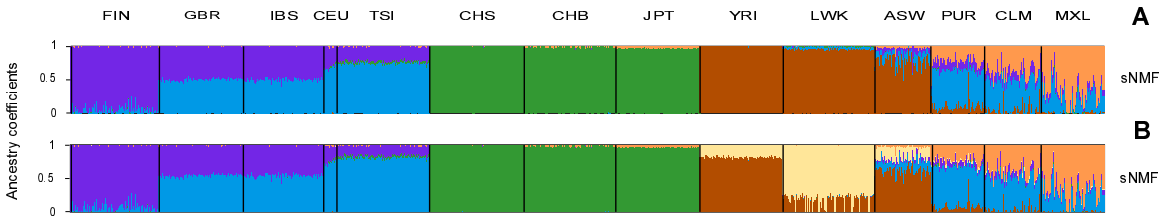

FIGURE 4.7: Graphical representation of ancestry estimates obtained for the 1000 Genomes Project data set. (A) Estimated ancestry coefficients using `sNMF` with $K = 5$ and $\alpha = 10,000$ (cross-entropy = 0.5010). (B) Estimated ancestry coefficients using `sNMF` with $K = 6$ and $\alpha = 10,000$ (cross-entropy = 0.5011) (FIN, Finnish ; GBR, British ; IBS, Spanish ; CEU, CEPH Utah residents ; TSI, Tuscan ; CHS, Southern Han Chinese ; CHB, Han Chinese ; JPT, Japanese ; YRI, Yoruba ; LWK, Luhya ; ASW, African-American ; PUR, Puerto Rican ; CLM, Colombian ; MXL, Mexican-American).

we displayed the Q matrices computed by each program for the Harvard HGPD panel HGDP00778 (78,000 SNPs), the HGDP-CEPH data (660,000 SNPs), the 1000 Genomes Project phase 1 data (2.2 million SNPs), and European populations of *A. thaliana* (216,000 SNPs). Using $K = 7$ ancestral populations for the Harvard HGPD panel HGDP00778, the cross-entropy criterion was 0.747 for the `ADMIXTURE` run, and it was 0.762 for the `sNMF` run. The criterion favored `ADMIXTURE` in this case, but the two runs led to very close estimates of the Q matrix ($R^2 = 0.99$, Figure 4.6). When the programs were applied to the HGDP-CEPH data ($K = 7$), the cross-entropy criterion was 0.691 for the `ADMIXTURE` run and 0.704 for `sNMF` (Figure 4.6). This particular `ADMIXTURE` run identified clusters that separated the African hunter–gatherer populations from the other populations, whereas `sNMF` identified a unique cluster in Africa. In the `sNMF` run, Middle East populations were separated from European populations (Figure 4.6). The differences between `ADMIXTURE` and `sNMF` results disappeared when additional runs were performed with distinct random seeds. Using $K = 5$ for the 1000 Genomes Project phase 1 data, `sNMF` identified clusters that correspond to the main geographic regions of the world, similarly to `ADMIXTURE` (Figure 4.7, cross-entropy = 0.5010). Substantial levels of European ancestry in African-Americans, Mexican-Americans, Puerto Ricans, and Colombians were inferred by `sNMF` and by the other program. An interesting case was with the application of ancestry estimation programs to European populations of *A. thaliana*, a selfing plant characterized by high levels of inbreeding (Atwell et al., 2010). Using $K = 3$, the cross-entropy criterion for `ADMIXTURE` was 0.641 on average, while the average value for `sNMF` was 0.483. The value of the criterion suggests that `sNMF` estimates were more accurate than those obtained from `ADMIXTURE`. The graphical output of the Q matrix displayed clinal variation of ancestry coefficients occurring along an East–West gradient separating two clusters, and Northern Swedish accessions were grouped into a separate cluster. These results supported previous estimates based on sequence data (Supplementary Figure 4.8) (François et al., 2008).

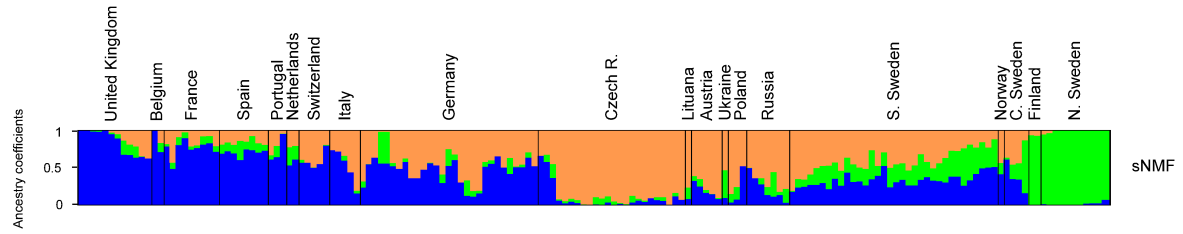

FIGURE 4.8: Graphical representation of admixture estimates for European populations of *A. thaliana*. Estimated admixture coefficientscients using `sNMF` using $K = 3$ and $\alpha = 100$ (cross-entropy = 0.483).

4.3.5 Simulated data analysis

To further ascertain the accuracy of `sNMF` estimates and to compare those estimates with `ADMIXTURE`, we employed computer simulations based on the 1000 Genomes Project and *A. thaliana* data sets. We also assessed the ability of the cross-entropy criterion to correctly identify the value of K when it is known.

TABLE 4.4: Statistical errors for `ADMIXTURE` and `sNMF` on simulated data sets

Ancestry estimation	Dir(1, 1, 1)	Dir(0.5, 0.5, 0.5)	Dir(0.1, 0.1, 0.1)	Dir(0.2, 0.2, 0.05)	Dir(0.2, 0.2, 0.5)	Dir(0.05, 0.05, 0.01)
Q matrix						
ADMIXTURE	0.023	0.012	0.004	0.006	0.010	0.003
sNMF $\alpha = 0$	0.020	0.011	0.007	0.007	0.013	0.009
sNMF $\alpha = 100$	0.024	0.014	0.006	0.006	0.014	0.006
G matrix						
ADMIXTURE	0.029	0.022	0.016	0.022	0.022	0.022
sNMF $\alpha = 0$	0.034	0.027	0.021	0.028	0.028	0.028
sNMF $\alpha = 100$	0.034	0.027	0.021	0.028	0.028	0.028

Dir: Dirichlet distribution used to simulate "true" admixture coefficients, using three ancestral populations.

In a first series of simulations, we used the 1000 Genomes Project data to generate genotypes showing various levels of admixture. As our ancestral populations, we chose the CHB, GBR, and YRI samples (1000 Genomes Project Consortium., 2012), and true Q matrices were created using several parameterizations of the Dirichlet distribution (Table 4.4). Genotypic matrices were simulated according to the binomial model used by `ADMIXTURE`. In this context, `ADMIXTURE` estimates are thus expected to be more accurate than `sNMF` estimates. For the range of parameters explored in the simulations, root mean-squared errors comparing the estimated and true values of the Q matrix remained $< 2\%$ for both programs (Table 4.4). For moderate levels of admixture, differences in statistical errors were $< 1\%$ regardless of the value of the regularization parameter, α, used in `sNMF`. This result indicated that `sNMF` estimates are generally accurate and that relatively small values of α do not influence `sNMF` outputs for data sets of size comparable to those used in simulations. Root mean-squared errors comparing the estimated and true values of the G matrix were slightly lower for `ADMIXTURE` than for `sNMF` (Table 4.4). This could be explained as we simulated from a binomial model (unrelated individuals) and as the

TABLE 4.5: Choice of K for `sNMF` using the cross-entropy criterion (simulated data).

	Dir(1, 1, 1)	Dir(.5, .5, .5)	Dir(.1, .1, .1)	Dir(.2, .2, .05)	Dir(.2, .2, .5)	Dir(.05, .05, .01)
$K = 2$	0.713	0.703	0.682	0.662	0.706	0.645
$K = 3$	**0.707**	**0.691**	**0.660**	**0.642**	**0.697**	**0.624**
$K = 4$	0.708	0.692	0.661	0.644	0.699	0.626
$K = 5$	0.710	0.694	0.663	0.645	0.700	0.628

Dir : Dirichlet distribution used to simulate "true" admixture coefficients using 3 ancestral populations.

number of degrees of freedom in `sNMF` is twice the number of degrees of freedom in `ADMIXTURE`. Regarding the choice of the number of clusters, the cross-entropy criterion was minimal for $K = 3$ for every simulated data set (Supplementary Table 4.5).

To evaluate the relative impact of linkage disequilibrium (LD) on `sNMF` and `ADMIXTURE` ancestry estimates, we considered subsets of SNPs sampled from the 1000 Genomes Project data set. We compared ancestry estimates computed by each program for data sets containing blocks of > 30 SNPs spaced < 20 kb apart and for data sets containing SNPs separated by > 20 kb (20,000 SNPs, 20 replicates). Using linked blocks of SNPs, the average value of the RMSE over all runs was 0.0859 (0.0297 for unlinked SNPs) for `sNMF` whereas it was 0.0976 for `ADMIXTURE` (0.257 for unlinked SNPs). Our results show that LD had an impact on the accuracy of ancestry estimates regardless of the program used and that the magnitude of the effect was similar for `ADMIXTURE` and `sNMF` (Supplementary Figure 4.9).

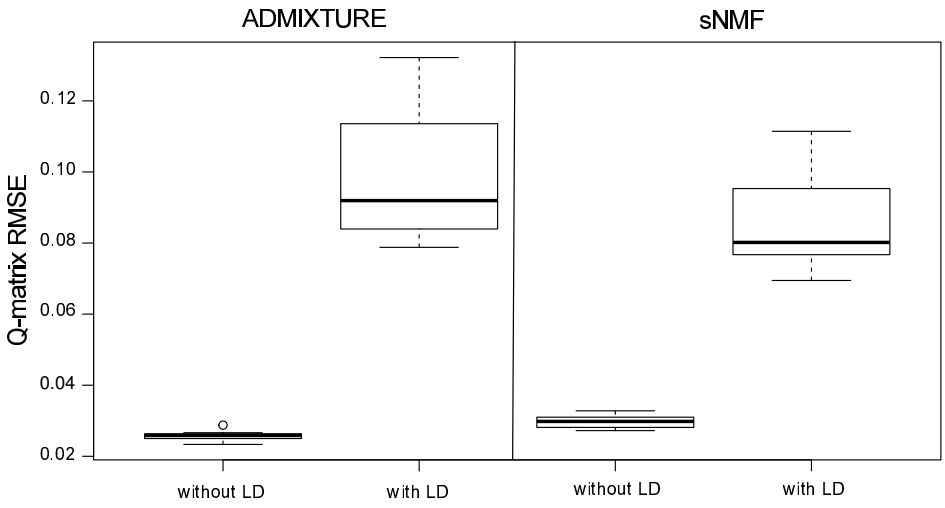

FIGURE 4.9: Accuracy of `ADMIXTURE` and `sNMF` in the presence of linkage disequilibrium. RMSEs between estimated Q matrices without and with linkage disequilibrium for `ADMIXTURE` and `sNMF` using $K = 5$ based on subsets of SNPs sampled from the 1000 Genome Project data set.

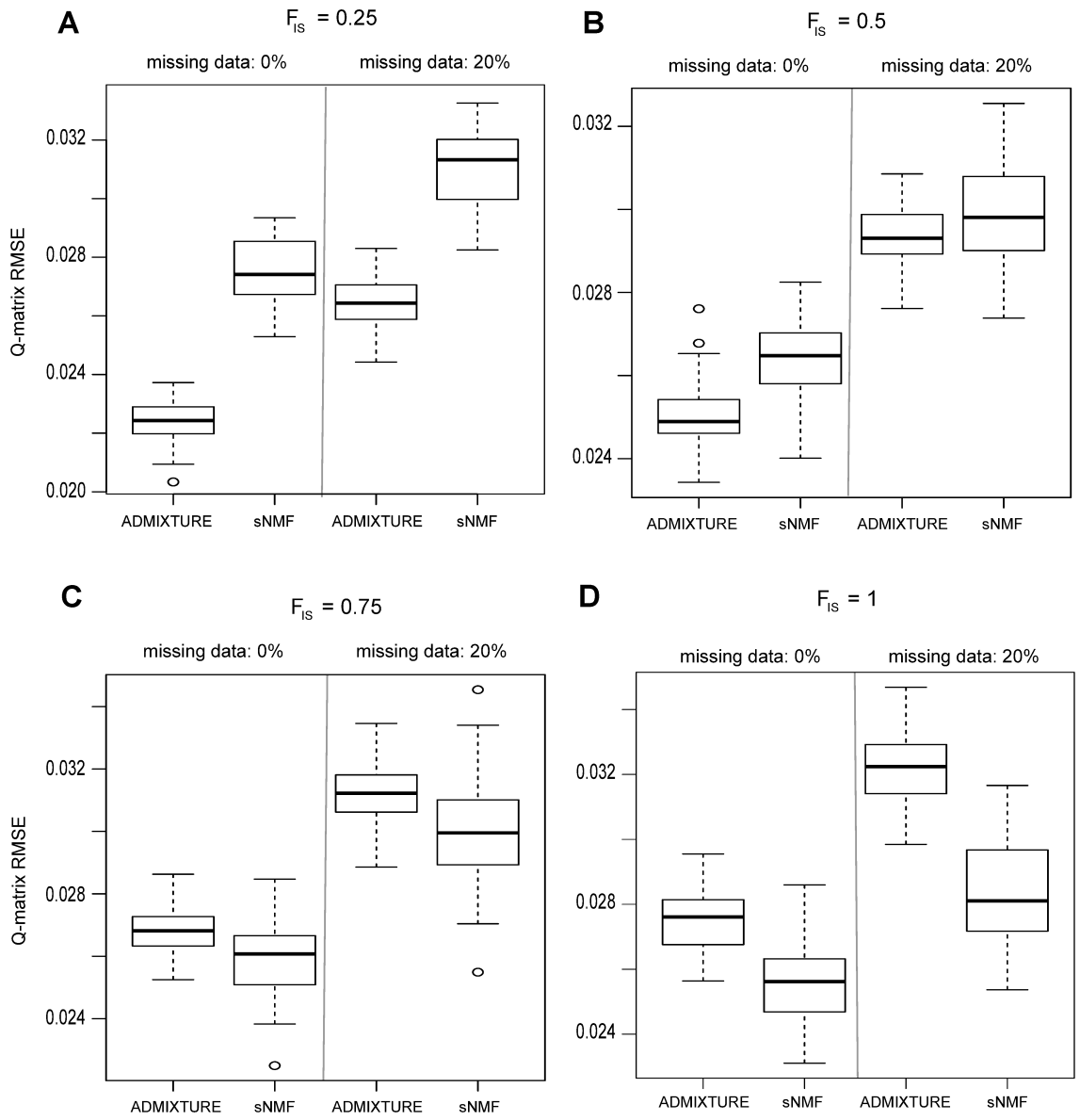

FIGURE 4.10: Accuracy of `ADMIXTURE` and `sNMF` in the presence of related individuals. Shown are RMSEs between estimated Q matrices and a known matrix used to generate simulated data. Simulations mimicked the population structure of European populations of *Arabidopsis thaliana*. (A and B) Moderate levels of inbreeding, F_{IS} = 25–50%. (C and D) Strong levels of inbreeding, F_{IS} = 75–100%.

We used another series of simulated data to evaluate the sensitivity of `ADMIXTURE` and `sNMF` estimates to the presence of related individuals and inbreeding in the sample. Based on empirical data, we used simulation models that mimicked the population structure of European populations of A. thaliana. First, we verified that the true value of the number of ancestral populations was correctly recovered by the `sNMF` program, using the cross-entropy criterion ($K = 3$). Next, we evaluated statistical errors for `ADMIXTURE` and `sNMF` estimates of the Q matrix. RMSEs remained $< 4\%$ for both programs. These

results showed that the two programs produced accurate estimates of the Q matrix in the presence of inbreeding and missing data (Figure 4.10).

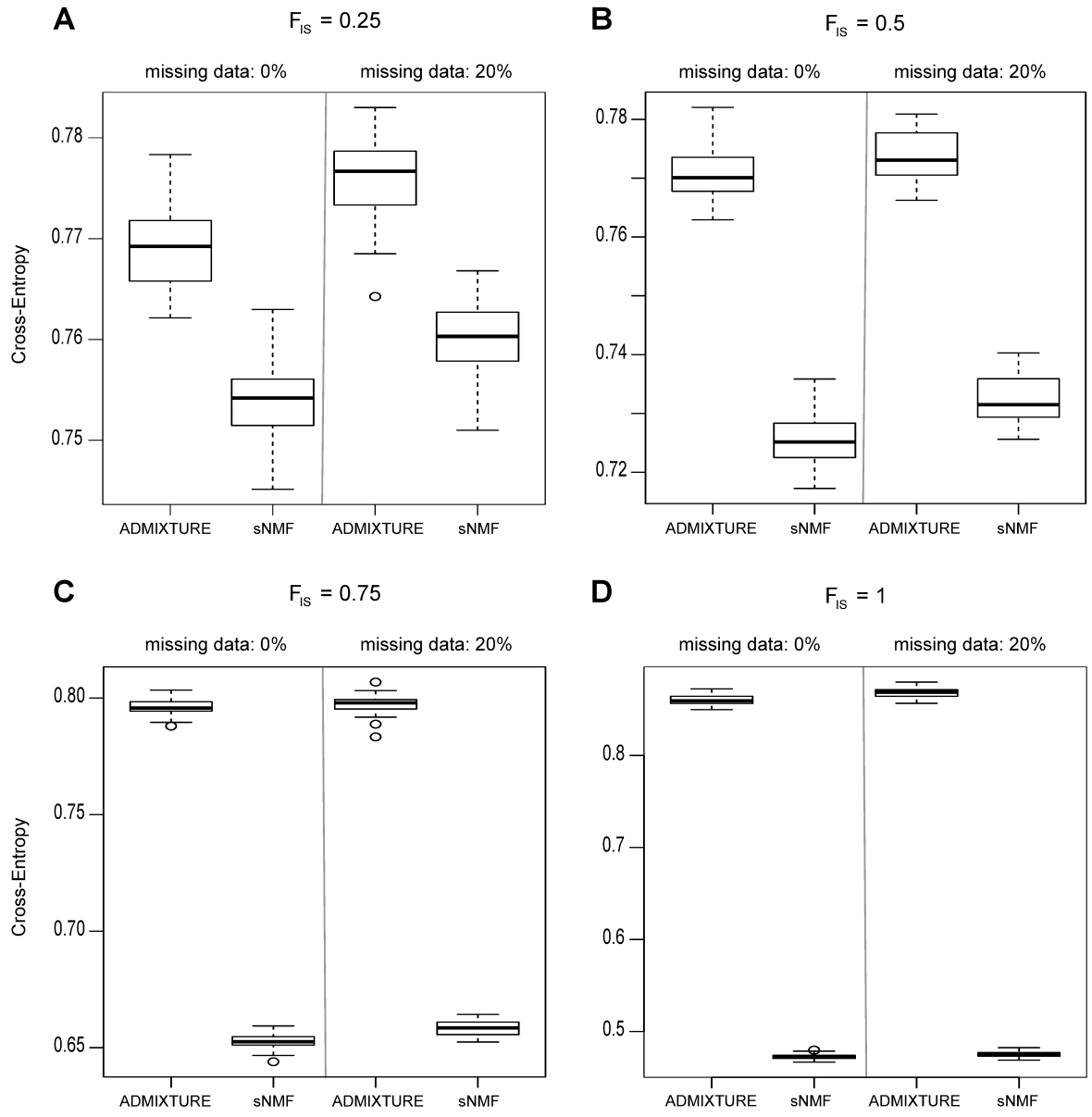

FIGURE 4.11: Accuracy of `ADMIXTURE` and `sNMF` in the presence of related individuals. Cross-entropy criterion for `ADMIXTURE` and for `sNMF`. Simulations mimicked the population structure of European populations of *Arabidopsis thaliana*. A-B) Moderate levels of inbreeding, $F_{IS} = 25 - 50\%$, C-D) Strong levels of inbreeding, $F_{IS} = 75 - 100\%$.

`ADMIXTURE` estimates were robust to the inclusion of moderate levels of inbreeding in the sample. When the values of the inbreeding coefficient were $0.25 - 0.5$, `ADMIXTURE` ancestry estimates were more accurate than `sNMF` estimates. When the values of the inbreeding coefficient were > 0.5 and for fully inbred lines, `sNMF` produced better estimates than `ADMIXTURE` (Figure 4.10). The cross-entropy criterion was smaller for `sNMF` than

for ADMIXTURE, showing that sNMF produced better prediction of masked genotypes than ADMIXTURE (Supplementary Figure 4.11). This result can be explained by a more accurate estimation of genotypic frequencies for sNMF than for ADMIXTURE in the presence of strong levels of inbreeding.

4.4 Discussion

We applied the computer program sNMF to the estimation of individual ancestry coefficients, using large population genetic data sets for humans and for *A. thaliana*, and compared the program performances to those of ADMIXTURE. For six HGDP data sets, ancestry estimates obtained with sNMF and ADMIXTURE strongly agreed with each other. In addition, the sNMF program was able to analyze the 1000 Genomes Project phase 1 data set within a few hours, using a standard computer processing unit. Without significant loss of accuracy, sNMF computed estimates of admixture proportions within runtimes that were $\approx 10 - 30$ times faster than those of ADMIXTURE.

The approach used by sNMF is based on theoretical connections between likelihood approaches, PCA, and NMF methods (Ding et al., 2008; Engelhardt and Stephens, 2010; Lawson et al., 2012; Parry and Wang, 2013). Several methods can be applied to computing NMF estimates, including the multiplicative update algorithm, the projected-gradient method, and the alternating least-squares algorithm (Brunet et al., 2004; Berry et al., 2007; Kim and Park, 2011). For population genetic data, we found that alternating least-squares algorithms coupled with the active set method provided the best trade-off between speed and accuracy and improved performance significantly over other NMF implementations (Kim and Park, 2011).

To decide which algorithm yielded the best estimates, we introduced a predictive criterion based on the computation of cross-entropy and the imputation of masked genotypes. For HGDP data sets, the cross-entropy criterion discarded large values of the sNMF regularization parameter (> 1000). For the large data sets, a wide range of values of the regularization parameter reached similar predictive values. For data sets having <10,000 SNPs, we found that parsimony (i.e., large values of α) could improve estimation of ancestry coefficients. We observed that a likelihood approach could benefit the analysis of modest-sized data sets or data containing a large number of missing genotypes. For larger data sets and missing $< 20\%$ genotypes, sNMF ancestry estimates were statistically close to those obtained with ADMIXTURE, and both programs were equally efficient at predicting masked genotypes. Statistical theory actually predicts that errors in evaluating the cross-entropy criterion are of order $O(1/\sqrt{n_L})$ where n_L is the number of masked genotypes.

For Harvard HGDP panels, differences between the `ADMIXTURE` and `sNMF` results could be considered hardly significant and estimates were statistically similar. The example of the Harvard HGDP panels showed that the cross-entropy criterion could also be used to discriminate among program runs regardless of the program used.

The assumptions underlying STRUCTURE and `ADMIXTURE` rely on simplified population genetic hypotheses. More specifically, the assumptions include absence of genetic drift and Hardy–Weinberg and linkage equilibrium in ancestral populations. The coding used by `sNMF` enabled the estimation of homozygote and heterozygote frequencies and avoided Hardy–Weinberg equilibrium assumptions. Although `ADMIXTURE` analyses were robust to small departures from Hardy–Weinberg equilibrium in human data, `sNMF` was more appropriate to deal with inbred lineages. For European populations of *A. thaliana*, the values of the cross-entropy criterion indicated better predictive results for `sNMF` than for `ADMIXTURE`. The difference between `sNMF` and `ADMIXTURE` predictions could be explained as the binomial model of `ADMIXTURE` is not suited to the high levels of inbreeding observed in *A. thaliana* populations (Atwell et al., 2010). As seen from Equation 1, an implicit assumption underlying NMF predictions is that genotypic frequencies can be formed according to instantaneous mixtures of ancestral frequencies without genetic drift. Interpretations of admixture using estimates obtained using likelihood and least-squares methods can be confounded by the existence of phylogenetic relationships among population samples (see Patterson et al. (2012) for an alternative approach) or by complex demographic scenarios such as spatial range expansion (François and Durand, 2010).

Comparing the relative computational performances of `ADMIXTURE` and `sNMF` was a difficult task because runtimes are dependent on several factors. These factors include the size and other characteristics of each data set, the tolerance threshold used when stopping program iterations, the use of multiprocessor algorithms, and the initial values of the Q and G matrices. For example, runtimes could be shortened by using initial values obtained after running the program on reduced data sets.

We explain the relative speed of the NMF algorithm by looking at algorithmic complexity for each program. The ANLS algorithm iterates cycles that solve linear regression equations for Q and G. The complexity of a single cycle of `sNMF` is of order $O(KLn)$, where K is the number of clusters, n the number of individuals, and L the number of loci. The complexity of a single cycle of `ADMIXTURE` is of order $O(K^2Ln)$ (Alexander et al., 2009). Since the default tolerance threshold in this program implies that the program generally runs a small number of cycles (e.g., < 40 cycles for the 78,000-SNPs Harvard HGDP panel), we observed that least-squares algorithms ran significantly faster than likelihood algorithms when analyzing large population genomic data sets with large values of K. The `sNMF` program can be downloaded from http ://membres-timc.imag.fr/Olivier.Francois/snmf.html.

4.5 Acknowledgments

We thank Nick Patterson, Eric Stone, Badr Benjelloun, and an anonymous reviewer for their useful comments on a previous version of this manuscript. We thank the 1000 Genomes Project for authorizing us to use the phase 1 data. This work was supported by a grant from la Région Rhône-Alpes to Eric Frichot and Olivier François. Olivier François acknowledges support from Grenoble Institute of Technology.

Chapitre 5

Tests d'associations entre des locus et des gradients environnementaux utilisant des modèles mixtes à facteurs latents

L'adaptation locale à des environnements agit, à travers la sélection naturelle, sur un grand nombre de locus, chacun ayant un effet phénotypique faible. Une manière de détecter ces locus est d'identifier des polymorphismes génétiques qui montrent de fortes corrélations avec des variables environnementales utilisées en tant qu'indicateurs de pressions écologiques. Dans cet article, nous proposons de nouveaux algorithmes fondés sur la génétique des populations, la modélisation en écologie et des techniques d'apprentissage statistique afin de cribler des génomes à la recherche de signatures d'adaptation locale. Implantés dans le programme `LFMM`, ces modèles mixtes à facteurs latents utilisent une approche dans laquelle la structure de population est modélisée par des variables non-observées. Ces algorithmes rapides et efficaces détectent des corrélations entre des variations génétiques et environnementales tout en estimant simultanément les différents niveaux de structure de population. Comparer ces algorithmes avec des méthodes similaires nous a montré que le logiciel `LFMM` peut estimer efficacement les effets aléatoires dus à l'histoire des populations et aux patrons d'isolement par la distance lorqu'il estime les corrélations entre genes et environnements. De plus, le logiciel `LFMM` diminue le nombre de fausses associations dans des études de criblage génomique. Nous avons appliqué, le logiciel `LFMM` à des jeux de données humaines et de plantes, en mettant en avant plusieurs gènes avec des fonctions associées au développement qui montrent de fortes corrélations avec des gradients climatiques.

Testing for associations between loci and environmental gradients using latent factor mixed models

Eric Frichot, Sean D Schoville, Guillaume Bouchard, Olivier François.

Molecular Biology and Evolution. 2013 ; 30 : 1687–1699.

Abstract

Adaptation to local environments often occurs through natural selection acting on a large number of loci, each having a weak phenotypic effect. One way to detect these loci is to identify genetic polymorphisms that exhibit high correlation with environmental variables used as proxies for ecological pressures. Here, we propose new algorithms based on population genetics, ecological modeling, and statistical learning techniques to screen genomes for signatures of local adaptation. Implemented in the computer program "latent factor mixed model" (LFMM), these algorithms employ an approach in which population structure is introduced using unobserved variables. These fast and computationally efficient algorithms detect correlations between environmental and genetic variation while simultaneously inferring background levels of population structure. Comparing these new algorithms with related methods provides evidence that LFMM can efficiently estimate random effects due to population history and isolation-by-distance patterns when computing gene-environment correlations, and decrease the number of false-positive associations in genome scans. We then apply these models to plant and human genetic data, identifying several genes with functions related to development that exhibit strong correlations with climatic gradients.

Keywords

local adaptation, environmental correlations, genome scans, latent factor models, population structure

5.1 Introduction

Local adaptation through natural selection plays a central role in shaping the variation of natural populations (Darwin, 1859; Williams, 1966) and is of fundamental importance in evolution, conservation, and global-change biology (Joost et al., 2007; Manel et al., 2010; Barrett and Hoekstra, 2011; Schoville et al., 2012; Jay et al., 2012). The intensity of natural selection commonly varies in space and can result in gene-environment interactions that have measurable effects on fitness (Storz and Wheat, 2010). This can lead to local adaptation if populations maintain locally advantageous traits despite gene flow with neighboring populations.

In principle, identifying chromosomal regions involved in adaptive divergence can be achieved by scanning genome-wide patterns of DNA polymorphism (Nielsen, 2005; Storz, 2005). Usually, the aim of screening procedures is to detect locus-specific signatures of positive selection. In populations inhabiting spatially distinct environments, loci that underlie adaptive divergence can be detected by comparing relative levels of differentiation among large samples of unlinked markers (Beaumont and Nichols, 1996; Beaumont and Balding, 2004) and by using empirical tests to compare levels of differentiation with the genomic background (Kelley et al., 2006; Akey, 2009; Novembre and Di Rienzo, 2009).

An alternative way to investigate signatures of local adaptation, especially when beneficial alleles have weak phenotypic effects, is by identifying polymorphisms that exhibit high correlation with environmental variables (Joost et al., 2007; Hancock et al., 2008; Poncet et al., 2010; Pritchard et al., 2010; Coop et al., 2010). In natural populations, quantitative traits that exhibit continuous geographic variation are often associated with specific ecological variables reflecting selective pressures acting on individual phenotypes (Endler, 1977). This type of variation is then reflected in geographic clines or in sympatric populations that exploit different ecological niches (Haldane, 1948; Berry and Kreitman, 1993; Young et al., 2005; Prugnolle et al., 2005). Evidence for local adaptation to continuous environments can be detected if there is highly significant association with the environmental variables at some loci compared with the background genomic variation.

A major difficulty is that the geographical basis of both environmental and genetic variation can confound interpretation of the tests (Eckert et al., 2010), as local adaptation can be hindered by gene flow (Lenormand, 2002), and can be difficult to distinguish from the effects of genetic drift and demographic history (Novembre and Di Rienzo, 2009). The main problem is that without corrections for the effect of population structure or isolation-by-distance (IBD), the underlying null distribution may be insufficient to account for the demographic history of the study organism. As a result, tests for associations between

loci and environmental variables using classical regression models will be prone to high rates of false positives (FP) (Meirmans, 2012). Recent studies have used the background patterns of allele frequencies to build a null model that accounts for the effects of drift and demographic history (Hancock et al., 2008; Coop et al., 2010; Fumagalli et al., 2011; Hancock et al., 2011). To correct for population stratification, Hancock et al. (2008) used an empirical approach that estimates the covariance of allele frequencies among populations. These authors assessed the evidence for local adaptation of each allele by testing whether environmental variables explained more variance than a null model with this particular covariance structure.

A drawback of empirical tests is the need to identify selectively neutral loci from the genomic background before testing for associations with environmental factors. The need to identify such a list a priori implies that tests based on empirical estimates of relatedness can lack power to reject neutrality, which is an important limitation for data sets where all loci are potentially under selection. For example, single nucleotide polymorphism (SNP) data sets derived from expressed sequences are often used to study local adaptation in nonmodel organisms (Eckert et al., 2010). Choosing a subset of markers not only reduces the size of such a data set but could also arbitrarily bias downstream statistical tests if only certain subsets of data (e.g., synonymous sites) are chosen as the neutral markers. After all, putatively neutral sites can be linked to loci under selection over large physical distances (Thibert-Plante and Hendry, 2010). In this study, we address this problem by introducing statistical models called latent factor mixed models (LFMM).

Using these models, we test correlations between environmental and genetic variation while estimating the effects of hidden factors that represent background residual levels of population structure. To perform parameter estimation, we extend probabilistic principal component analysis (PCA) and recent statistical learning approaches (Tipping and Bishop, 1999; Salakhutdinov and Mnih, 2008; Engelhardt and Stephens, 2010; Frichot et al., 2012). Based on low rank approximation of the residual covariance matrix, we implement algorithms to deal with hundreds of thousands of polymorphisms with rapid computing times. We show that our algorithms control for random effects due to population history and spatial autocorrelation when estimating gene-environment association, and we provide examples of how our approach can be used to detect local adaptation in plants and humans.

5.2 New Approaches

Consider the data matrix, $(G_{i\ell})$, where each entry records the allele frequency for individual i at the genomic locus ℓ, $1 \leq i \leq n$, $1 \leq \ell \leq L$, and n and L represent the total sample size and number of loci, respectively. For simplicity, we assume our loci are bi-allelic, for example, SNPs. In this case, for each marker, there is an ancestral and a derived allele, and $G_{i\ell}$ is the number of derived alleles for locus ℓ and individual i. For diploid data, $G_{i\ell}$ is thus equal to 0, 1, or 2, and corresponds to the genotype at locus ℓ. In addition to the genotypic data, we have a vector of d geographic and environmental variables, (X_i), for each individual. The vector of covariates could include latitude and longitude, habitat and other ecological information, climatic variables, and so forth, which serve as proxies for unknown environmental pressures (Hancock et al., 2008; Eckert et al., 2010).

5.2.1 Model

To evaluate associations between allele frequencies and environmental variables while correcting for background levels of population structure, we regard the matrix G as being a response variable in a regression mixed model

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell}\,, \tag{5.1}$$

where μ_ℓ is a locus specific effect, β_ℓ is a d-dimensional vector of regression coefficients, U_i and V_ℓ are scalar vectors with K dimensions that model latent factors and their scores $(1 \leq K \leq n)$. The residuals $\epsilon_{i\ell}$ are statistically independent Gaussian variables of mean zero and variance σ^2.

We refer to the earlier-mentioned statistical model as a LFMM (see Materials and Methods, section 5.5). Similar models, termed factor regression models, have been considered earlier in biostatistics in the inference of molecular pathways from gene expression data (West, 2003; Carvalho et al., 2008).

In LFMMs, environmental variables are introduced as fixed effects while population structure is modeled using latent factors. In the model, the matrix term $U^T V$ models the part of genetic variation that cannot be explained by the environmental pressures. Note that the use of factorization methods is closely related to estimating population structure by singular value decomposition, a well-established technique for identifying scores and loadings in PCA (Jolliffe, 1986). Recently, matrix factorization methods have been generalized to

include probabilistic PCA (Tipping and Bishop, 1999) and probabilistic matrix factorization algorithms (Salakhutdinov and Mnih, 2008), which have proven useful in analyzing population genetic data (Engelhardt and Stephens, 2010). To clarify the connection between LFMM and PCA, assume that no environmental variable is available. In this case, we set $\beta_\ell = 0$ for all locus ℓ. In matrix factorization algorithms, a data matrix G with n rows and L columns can be decomposed into a product of two matrices U and V, where U has n rows and K columns, and V is a $K \times L$ matrix. Following Patterson et al. (2006), we assume that the genotypic data are centered. We consider the matrix $Y_{i\ell} = G_{i\ell} - \bar{G}_{.\ell}$, where we have subtracted the mean value of each column, $\bar{G}_{.\ell} = \sum_{i=1}^{n} G_{i\ell}/n$. For each individual i and locus ℓ, the decomposition is as follows :

$$Y_{i\ell} = U_i^T V_\ell = \sum_{k=1}^{K} U_{ik} V_{k\ell} \,. \tag{5.2}$$

To estimate the factor vectors U_i and V_ℓ, the squared error is minimized on the set of observed data

$$\min_{U,V} \sum_{i,\ell} \left(Y_{i\ell} - \sum_{k}^{K} U_{ik} V_{k\ell} \right)^2 . \tag{5.3}$$

With $K = L$, this approach is similar to computing PCA loadings and scores (Jolliffe, 1986). The number of components K can, however, be chosen much lower than the number of loci or individuals. In simulations, we based our choice of K on Tracy–Widom theory (Patterson et al., 2006). In real applications, this choice of K may be replaced by estimates of population genetic structure obtained with clustering algorithms like `STRUCTURE` (Pritchard et al., 2000a) or `TESS` (Chen et al., 2007). When values of K are low our algorithm is essentially a low-rank approximation of the covariance structure (Eckart and Young, 1936), which leads to computationally fast estimation algorithms. To estimate the LFMM parameters, we implemented a Gibbs sampler algorithm (Materials and Methods, section 5.5 and Supplementary File S1, section 5.6). We computed $|z|$-scores for all environmental effects, and we tested the significance of these effects using the standard Gaussian distribution and Bonferroni correction for multiple testing.

Incorporating population genetic structure using estimates of principal components or ancestry coefficients is common in genome-wide association studies (Price et al., 2006; Yu et al., 2006; Zhou and Stephens, 2012), and in tests based on empirical approaches (Coop et al., 2010; Poncet et al., 2010). In this paragraph, we explain the distinction between LFMM and tests based on empirical covariance matrices. Suppose that we start

by computing PCA scores from the matrix Y for all individuals, and denote by $\tilde{U}_i$ the PCA scores for individual i. The product matrix $\tilde{U}\tilde{U}^T$ is thus equal to the empirical covariance matrix

$$\tilde{U}\tilde{U}^T = YY^T/n\,. \tag{5.4}$$

Now using the scores as covariates in a Bayesian regression model, we obtain

$$G = \mu + \beta^T X + \tilde{U}^T V + \epsilon. \tag{5.5}$$

By a change of variables, this is equivalent to fitting the model

$$G = \mu + \beta^T X + \tilde{\epsilon} \tag{5.6}$$

where the distribution of $\tilde{\epsilon}$ is a multivariate Gaussian distribution of the covariance matrix equal to $\sigma^2\mathrm{Id} + \sigma_V^2 YY^T/n$. Here, Id is the n-dimensional identity matrix, and σ_V^2 is the variance of factor coordinates. Setting $\sigma_V^2 = 1$ and considering small values of the scaling parameter σ^2, the model defined in equation (5.6) is nearly equivalent to the model implemented in empirical approaches. In a Bayesian Gaussian regression framework, incorporating PCA scores as covariates in an association model is equivalent to modeling residuals as Gaussian vectors with covariance depending on the empirical covariance matrix of the genotypic data. Thus, a major difference between methods is that the factor matrix U and the regression coefficients β are estimated by a two-stage procedure in empirical approaches, whereas it requires a single step in LFMMs.

5.3 Results

We designed experiments based on simulated data to answer the following questions : 1) Are tests based on LFMMs conservative or liberal ? 2) How does the LFMM algorithm perform compared with existing methods such as logistic or standard regression models (Joost et al., 2007), principal component regression model (PCRM), partial Mantel tests (PMTs) (Fumagalli et al., 2011), standard linear mixed models (Zhou and Stephens, 2012), and Bayesian mixed models (Coop et al., 2010) ?

5.3.1 Distribution of *P*-values under the Null Hypothesis.

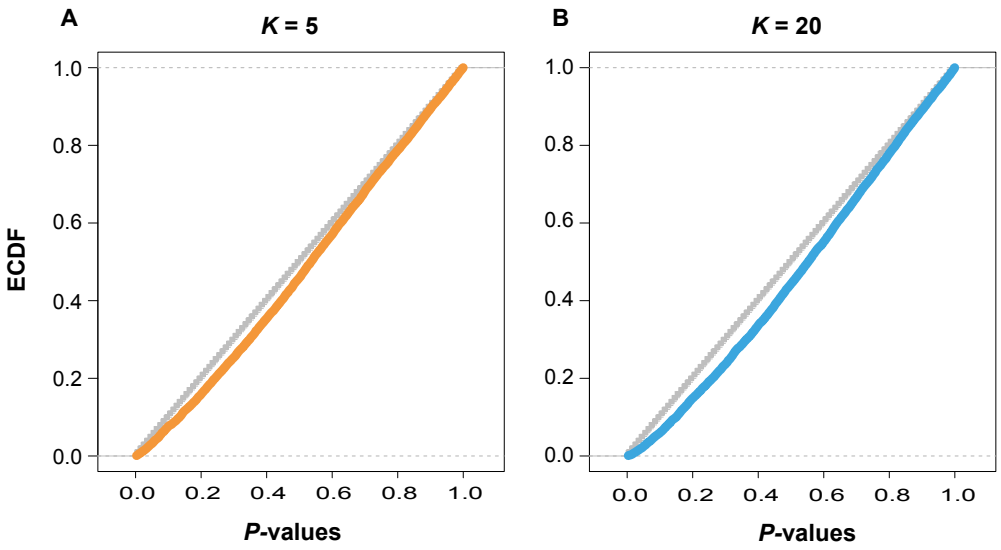

FIGURE 5.1: Simulations from the null model. Empirical cumulative distribution function (ECDF) of P-values for LFMM tests for simulations from a latent factor model using A) K = 5, B) K = 20 latent factors.

To evaluate the calibration of *P*-values, we used equation (1) with $\beta = 0$ to simulate data under a null hypothesis of no association with any environmental variable (Materials and Methods, section 5.5). Figure 5.1 reports the empirical cumulative distribution function (ECDF) of *P*-values for $K = 5$ and $K = 20$. Plots for other values of K are shown in Supplementary Figure 5.2. *P*-values are well calibrated when their ECDF is close to the uniform distribution, represented by the bisector line. Below the line, the test is conservative. For values of K less than 5, the ECDF was close to a uniform distribution, and *P*-values were correctly calibrated. For $K = 20$, the tests were slightly conservative. Thus, for moderate and for large values of the number of latent factors, the tests produced small numbers of FP associations.

Next, we used equation (1) to simulate data exhibiting various levels of population structure and association with a randomly generated environmental variable, and we compared the distributions of statistical errors for the following three estimation approaches : 1) LFMM, 2) a standard linear regression model, and 3) a PC regression model (Materials and Methods, section 5.5).

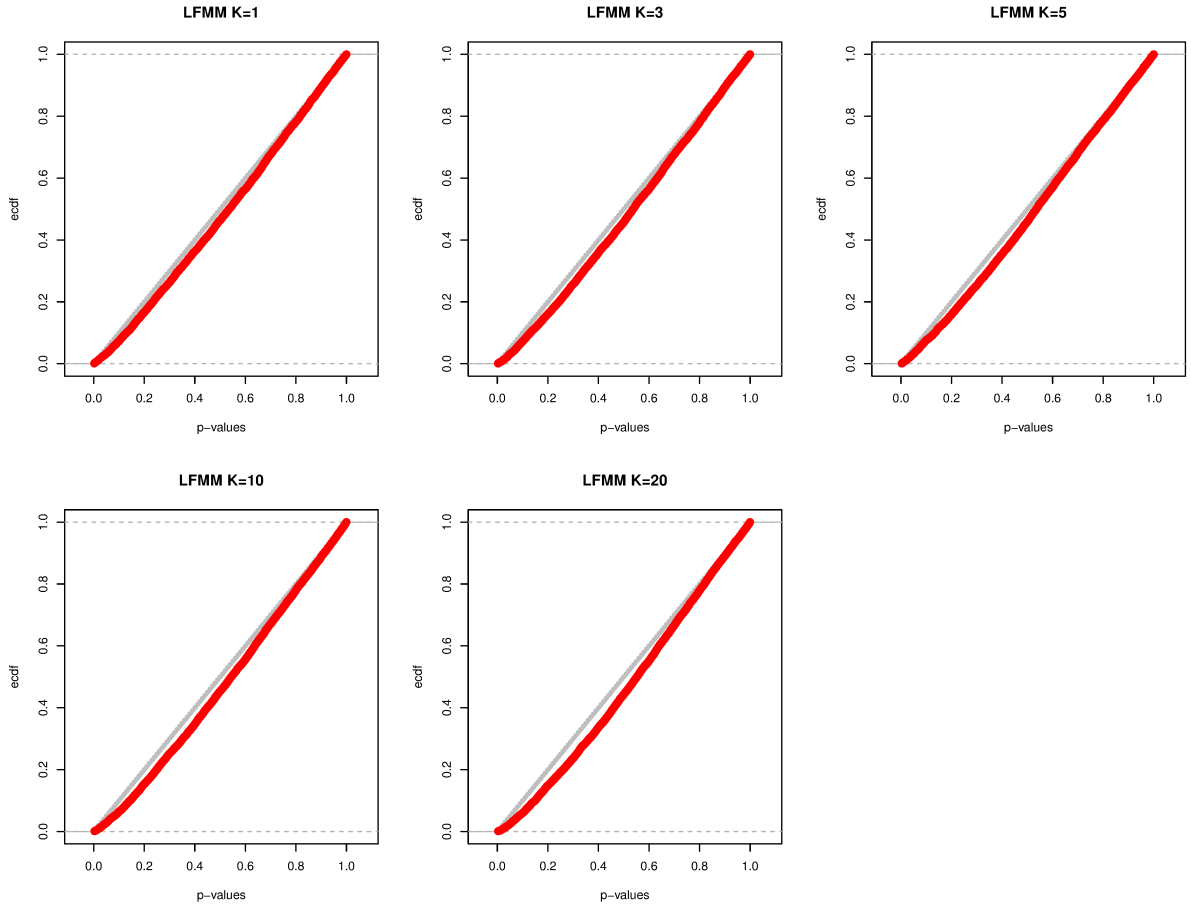

FIGURE 5.2: Simulations from the null model. Empirical cumulative distribution function for LFMM tests for simulations from generative models with $K = 1, 3, 5, 10$ and 20 latent factors.

TABLE 5.1: Mean squared errors for estimates of environmental effects.

K	LM	PCRM	LFMM
2	0.20	0.21	0.15
20	1.27	1.42	0.08
100	6.13	12.41	0.20

Figure 5.3 reports the quantiles of absolute errors for LFMM, the standard linear regression, and PC regression models. For LFMM, absolute errors ranged between 0 and 0.6 for $K = 2 - 20$, and between 0 and 1.0 for $K = 100$. Mean squared errors indicated that the bias and variance of estimates were small (Table 5.1). Compared with LFMMs, the relative errors of the linear and PC regression estimates increased with the rank of the hidden factor matrix. The absolute errors of these algorithms ranged between 0 and 1.4 for $K = 2$, between 0 and 3.2 for $K = 20$, and between 0 and 9.2 for $K = 100$. When linear or PC regression models were fitted to the data, the quantiles of errors shifted to values $\approx$ 1.74-fold higher for $K = 2$, $\approx$ 3.8 to 4.1-fold higher for $K = 20$, and $\approx$ 5.5 to

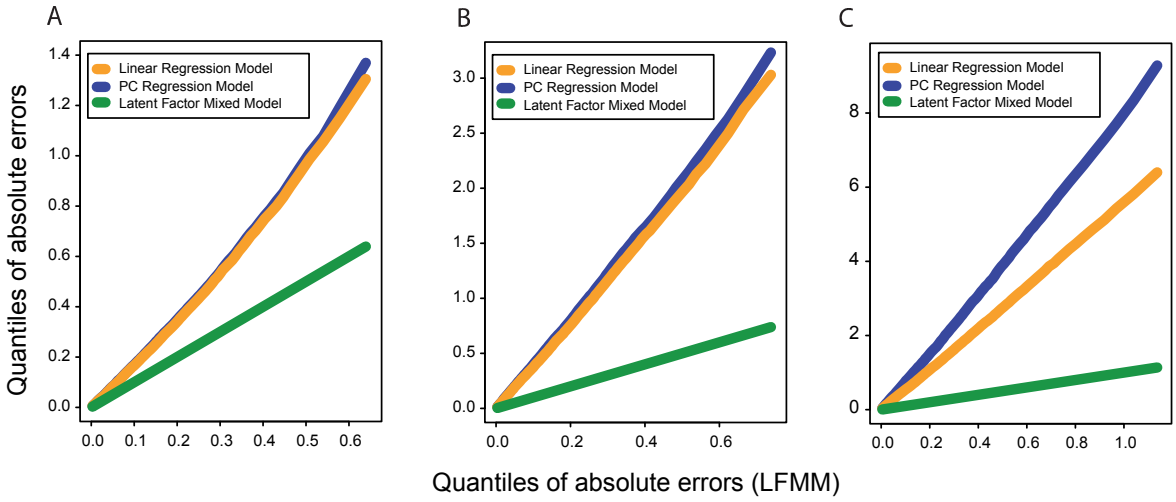

FIGURE 5.3: Generative model simulations. Quantiles of absolute errors for the standard linear regression, PC regression and LFM models using simulations from latent factor models using A) K=2, B) K=20 and C) K=100 latent factors.

7.7-fold higher for $K = 100$. Mean squared errors provided additional evidence of relatively poor performances of the linear regression and PC regression estimates when the levels of underlying structure increased (Table 5.1).

5.3.2 Spatial Coalescent Simulations

In another series of experiments, we compared the LFMM estimation algorithm against two methods that do not correct for population stratification, and against methods that use the empirical covariance matrix to correct for population stratification. The first set of methods include a linear model (LM) and generalized linear model (GLM) (Joost et al., 2007), and the second set of methods included three empirical methods : a PCRM, PMTs (Smouse et al., 1986; Legendre and Legendre, 2012), and the mixed models implemented in `BAYENV` and `GEMMA` (Coop et al., 2010; Zhou and Stephens, 2012). In PMTs, the relationship between population genetic distances at each SNP and a matrix of environmental variable distance was evaluated using a correction for correlations in genome-wide allele frequencies. With `GEMMA`, we implemented a standard linear mixed model in which a single environmental variable is explained by SNP genotype, and where relatedness is introduced by a random effect (see Materials and Methods, section 5.5, for a description of all methods).

To examine the outcome of tests when genetic variation is neutral at all loci, we computed the distributions of P-values under LM, GLM, PMT, PCRM, `GEMMA`, and LFMM with

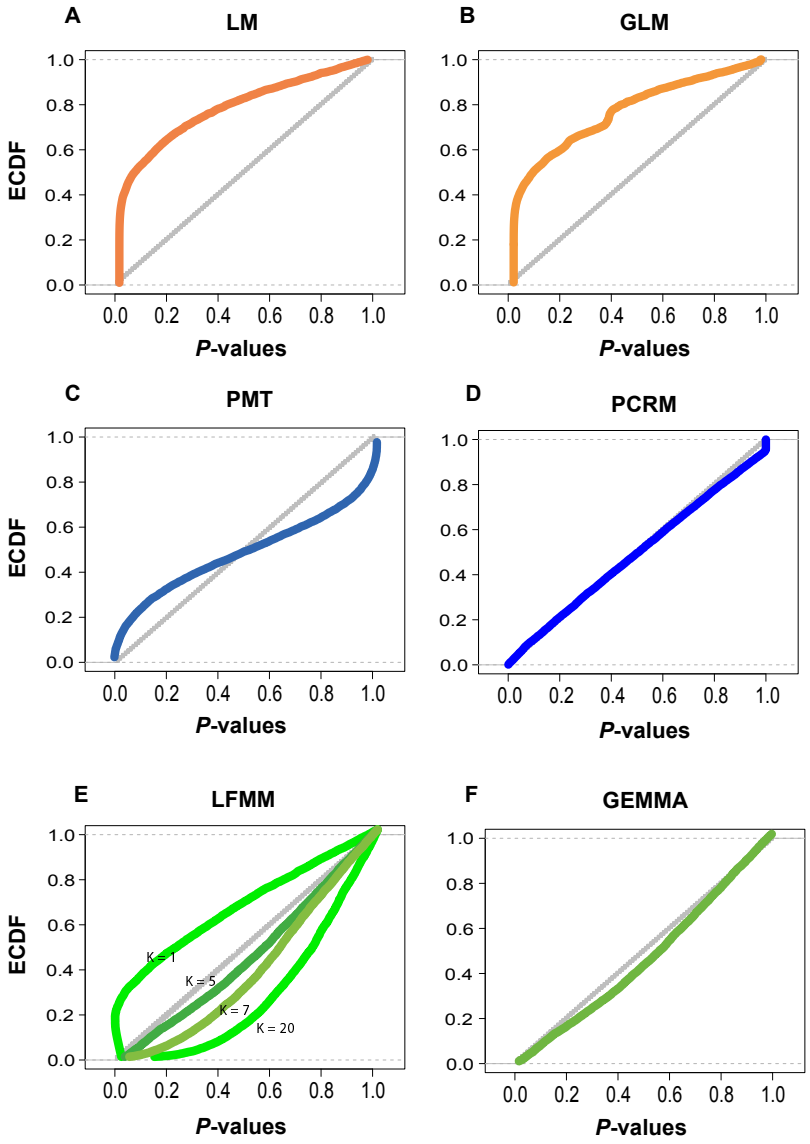

FIGURE 5.4: Spatial neutral coalescent simulations. Empirical cumulative distribution function (ECDF) of P-values for A) the linear regression model (LM), B) the generalized linear model (GLM), C) partial Mantel tests using Nei's genetic distance and the empirical correlation matrix for correction, D) the PC regression model using $K = 7$ principal components (PCRM), E) the LFM model using $K = 1, 5, 7, 20$ latent factors (LFMM) where the value $K = 5$ corresponds to the estimate of the number of clusters obtained from Bayesian clustering algorithms, and the value $K = 7$ is a Tracy-Widom estimate, and F) the standard linear mixed model implemented in `GEMMA`.

different values for the number of latent factors (K ranging from 1 to 20). The distributions of P-values for tests based on LM and GLM showed a strong departure from the uniform distribution (Figures 5.4A and 5.4B). In those cases, the tests were too liberal and produced a large number of FP results. For GLM, using population allele frequencies instead of individual genotypes reduced the number of FP associations but the tests based on these models remained liberal. The ECDF for PMTs showed an excess of low and

high P-values, but the curve was closer to a uniform distribution than with LM tests (Figure 5.4C). Using $K = 7$ PCs in PCRM, P-values for those tests and for `GEMMA` were well calibrated (Figures 5.4D-F). Choosing K based on Tracy–Widom theory ($K = 7$) and on Bayesian clustering algorithms ($K = 5$) led to slightly conservative tests for LFMMs (Figure 5.4E). ECDFs for all values of K are shown in Supplementary Figures 5.5 and 5.6, respectively.

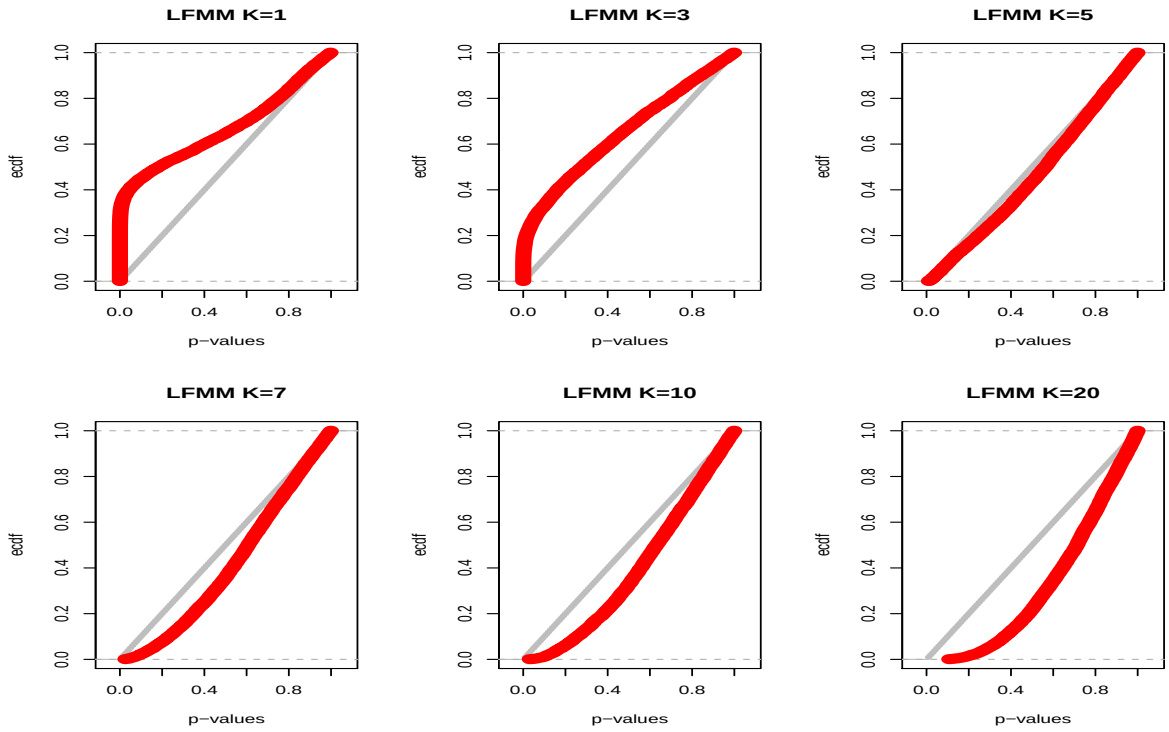

FIGURE 5.5: Spatial neutral coalescent simulations. Empirical cumulative distribution function for the LFMM using $K = 1, 3, 5, 10$ and 20 latent factors.

Next, we evaluated the ability of LFMMs to detect loci exhibiting correlations with particular environmental gradients and compared tests based on LFMMs with methods based on linear models. An environmental variable, x, was defined for each population as the geographic identifier of the population in the linear stepping-stone model. Following Haldane (1948), we chose a sigmoid function to represent the shape of a selected allele frequency cline through geographic space. Under strong selection ($\theta = 0.2$), we expect that tests produce low rates of FP associations while still preserving reasonable power.

For all simulated data sets, we evaluated the rates of false negative (FN) and of FP tests based on LM, GLM, PMT, PCRM, `GEMMA`, and LFMM for two values of the type I error (Table 5.2). In the case of strong selection, we found that standard linear models exhibited high rates of FP. In contrast, tests that include corrections for population structure—based on PMTs, PCRM, and standard linear mixed models—exhibited low rates of FP. But PMT, PCRM, and `GEMMA` exhibited large rates of FN, and these tests had

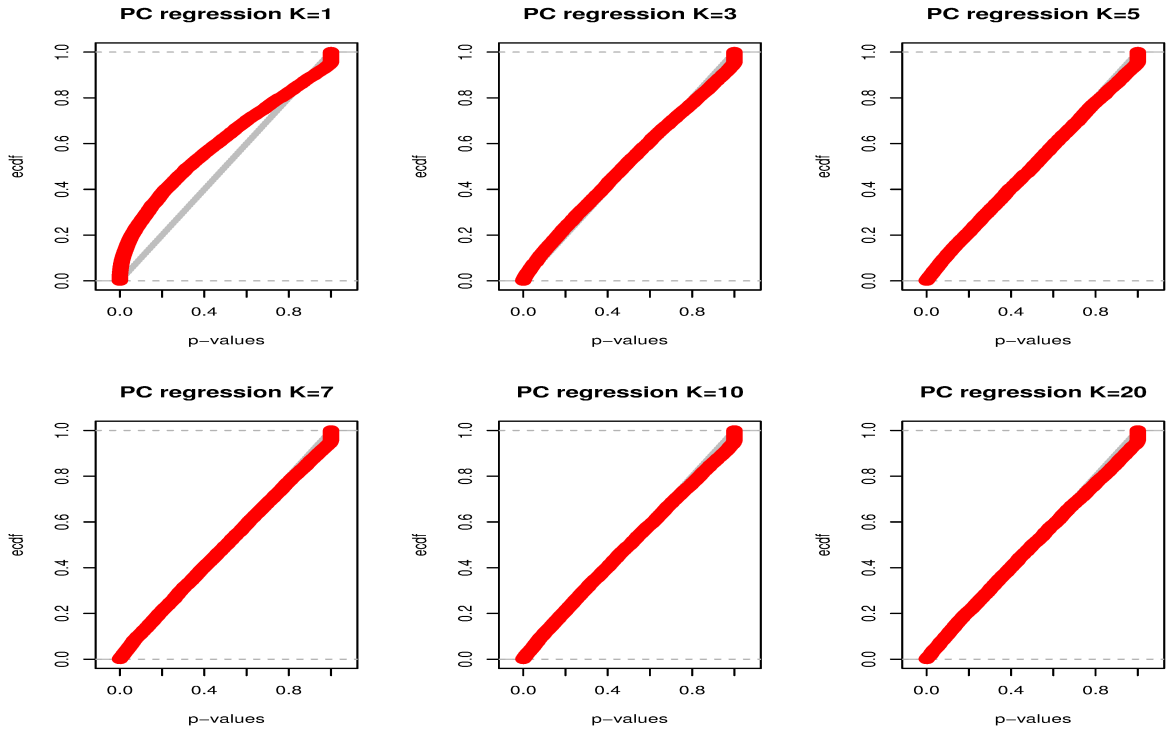

FIGURE 5.6: Spatial neutral coalescent simulations. Empirical cumulative distribution function for the PC regression model using $K = 1, 3, 5, 10$ and 20 latent factors.

low power to reject neutrality. These results provide evidence that the standard methods may fail to identify selected loci from the genomic background even though association with the environment is strong. In this context, tests based on LFMM produced low rates of FP and had reasonable power to reject neutrality (Table 5.2).

TABLE 5.2: Table 2 : Rates of false negative (FN) and false positive (FP) association for tests based on linear models (LM), principal component regression (PCRM), standard linear mixed models (`GEMMA`), Partial Mantel correlations (PMT) and LFM models (LFMM).

FN (FP)	LM	GLM	PCRM	GEMMA	PMT	LFMM
Type I error						
$\alpha = 0.001$	0% (33%)	0% (24%)	100% (3%)	100% (2%)	99% (6.8%)	4% (5%)
$\alpha = 0.0001$	0% (27%)	0% (19%)	100% (0%)	100% (0%)	100% (3.4%)	14% (3%)

To perform comparisons with the program `BAYENV` (Coop et al., 2010), we wanted to evaluate whether the program was able to detect weak selection. Thus, we set the intensity of selection through space to a low level ($\theta = 0.1$, Materials and Methods, section 5.5). As `BAYENV` returns Bayes factors instead of P-values, we considered ranked lists recording the M loci corresponding to the strongest associations (M between 1 and $L = 1,050$). Figure 5.7 reports the number of true positives (TP) as a function of the number of FP. Considering the rates of TP and FP, the mean area under the receiver-operating characteristic curve (AUC) for tests based on LFMMs with $K = 5 - 7$ factors were approximately $0.95 - 0.96$, whereas the AUC for `BAYENV` was equal to 0.88. In the linear stepping stone

model simulations, the tests based on LFMM performed better than BAYENV for all values of K (Figure 5.7).

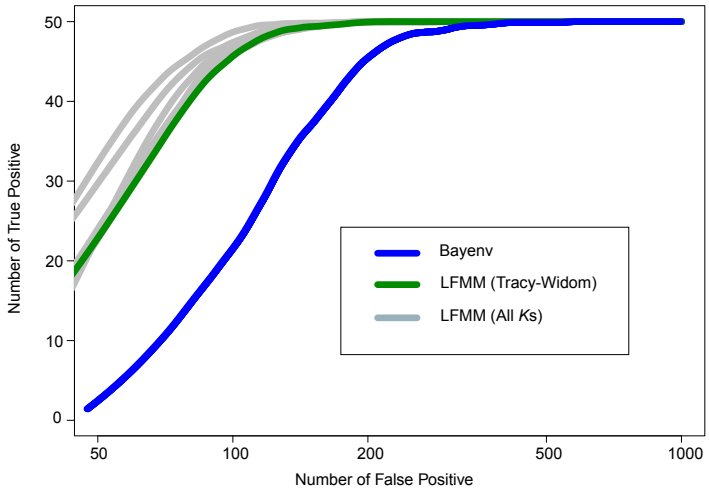

FIGURE 5.7: Spatial coalescent simulations with loci under selection. Number of true positive associations for BAYENV and for LFMM for $K = 5, 7$ (STRUCTURE and Tracy-Widom values) and for $K = 1, 3, 10, 20$ for spatial coalescent simulations including 1050 loci with 50 SNPs under selection.

5.3.3 Loblolly Pine

To illustrate the application of LFMMs, we analyzed genomic data of Loblolly pines (*Pinus taeda*, Pinaceae, Eckert et al. (2010). The Loblolly pine is distributed throughout the Southeastern United States, ranging from the arid Great Plains to the humid Eastern Temperate Forest ecoregion. These data consisted of 1,730 SNPs selected in expressed sequence tags (ESTs) for 682 individuals (Eckert et al., 2010).

We applied LFMM to the Loblolly pine data, testing 5 environmental variables representing the 5 first components of a PCA for 60 climatic variables (data from Eckert et al. (2010)). A total of 392, 113, and 30 SNPs obtained $|z|$-scores greater than 3, 4, or 5 for at least one environmental variable, respectively. On the basis this result, we considered that a SNP effect was significant when its $|z|$-score was greater than 4 (two-sided test). The cutoff $|z| > 4$ corresponds to P-values $P < 10^{-5}$ obtained after applying a Bonferroni correction for a type I error $\alpha = 0.01$ and $L \approx 10^3$ loci. Among the 50 loci with the highest $|z|$-scores, 17 were shared with those detected by Eckert et al. (2010) using BAYENV. Seven of the 10 SNPs with Bayes factors greater than 10^3 were confirmed by the LFMM analysis. For the first and second environmental variables, the two SNPs which obtained the highest Bayes factors using BAYENV were recovered by the LFMM analysis.

Table 5.3 provides a list of SNPs associated with climatic gradients and their functional annotation. The LFMM analysis discovered new significant and interesting associations with climatic gradients not identified in the analysis of Eckert et al. (2010), such as the chloroplast lumen 19 kDA protein involved in photosynthesis ($|z| = 6.42$), a pentatricopeptide repeat protein involved in oxidative stress and salt stress ($|z| = 5.90$), and the heat shock transcription factor hsf5 ($|z| = 5.60$) involved in regulation of transcription and response to temperature stress (Table 5.3 and Supplementary Table 5.4).

TABLE 5.3: Loblolly Pines. Annotation and gene ontology for some interesting SNPs with z-scores with absolute value greater than 4 for the first two components of 60 climatic variables.

Annotation	Gene Ontology	$-\text{Log}_{10}(P \text{ Value})$
Thylakoid lumenal 19 kDa chloroplast	Oxygen-evolving complex; Photosystem II	9.87
Pentatricopeptide repeat protein	Oxidative stress; salt stress	8.44
Conserved hypothetical protein	Ubiquitin-specific protease	8.28
Chalcone synthase	Flavonoid biosynthesis; wound response; oxidative stress	7.80
Heat shock	Temperature stress	7.67
Dirigent protein pdir18	Disease response	6.56
Heat shock transcription factor hsf5	Regulation of transcription; response to stress	6.15
Zinc finger	Transcription; DNA binding; zinc ion binding	5.84
Probable *n*-acetyltransferase hookless 1	Auxin signaling; photomorphogenesis; ethylene response	5.78
Calcium-binding pollen allergen	Polcalcin; calcium ion binding	4.61
Geranylgeranyl diphosphate synthase	Cholesterol biosynthesis; isoprenoid biosynthesis	4.59
Hypothetical protein OsI_04393	Trehalose-6-phosphate phosphatase	4.59
Potassium proton antiporter	Potassium ion transport; solute:hydrogen antiporter	5.54
DNA mismatch repair	DNA repair; regulation of DNA recombination	5.44

NOTE.—Annotation and gene ontology for some interesting SNPs with z-scores with absolute value greater than 4 for the first two components of 60 climatic variables.

TABLE 5.4: Loblolly pines. SNP identifier and annotation for SNPs with z-scores with absolute value greater than 4 for the first two components of 60 climatic variables.

SNP	Annotation	$-\log_{10}$(P-value)
2-4107-01-438	thylakoid lumenal 19 kda chloroplast	9.87
0-10719-01-95	pentatricopeptide repeat protein	8.44
2-1087-01-86	conserved hypothetical protein [Ricinus communis]	8.28
2-1818-01-168	chalcone synthase	7.80
CL17Contig1-03-443	heat shock	7.67
0-9449-02-292	dirigent protein pdir18	6.56
0-18317-01-495	potassium proton antiporter	6.46
UMN-CL194Contig1-04-130	dna mismatch repair	6.24
0-17238-01-294	Nodulin MtN21 family protein	6.20
0-17776-01-96	heat shock transcription factor hsf5	6.15
0-8823-01-306	squamosa promoter-binding	5.91
2-4856-01-162	zinc finger	5.84
0-4838-01-307	probable n-acetyltransferase hookless 1	5.78
2-3236-01-225	arabinogalactan-like protein	5.76
0-768-02-400	protein kinase family protein	5.72
UMN-5299-01-201	importin-alpha re-	5.33
2-4724-01-136	protein kinase	4.98
0-18887-02-633	amino acid transporter	4.94
CL996Contig1-03-68	af448201 1 alpha-xylosidase	4.93
2-3884-02-413	sf21d1 splice variant protein	4.92
UMN-CL148Contig1-02-220	Histone 2	4.82
CL3851Contig1-05-68	proliferating cell nuclear antigen	4.81
CL2121Contig1-05-658	glycolipid transfer	4.74
CL763Contig1-06-141	calcium-binding pollen Polcalcin	4.61
0-16664-01-58	geranylgeranyl diphosphate synthase	4.59
UMN-1598-02-647	hypothetical protein OsI 04393 [Oryza sativa Indica Group]	4.59
2-7619-01-193	target of myb1	4.57
2-2125-01-274	nodulation receptor kinase	4.40
CL3162Contig1-02-257	small gtp-binding protein	4.32
0-13722-01-343	dirigent-like protein	4.22
0-8922-01-655	TIFY domain containing protein	6.01
0-18317-01-495	potassium proton antiporter	5.54
UMN-CL194Contig1-04-130	dna mismatch repair	5.44
CL1381Contig1-01-188	aintegumenta-like protein	5.26
CL3851Contig1-05-68	proliferating cell nuclear antigen	4.52
2-2125-01-274	nodulation receptor kinase	4.27

5.3.4 Human data analysis

We applied LFMM to a worldwide sample of genomic DNA from 1,043 individuals in 52 populations, referred to as the Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain (HGDP–CEPH) Human Genome Diversity Cell Line Panel (hagsc.org/hgdp/). We extracted climatic data for each of the 52 population samples using the WorldClim data set at 30 arcsecond ($1km^2$) resolution (Hijmans et al., 2005) (Supplementary Table 5.5).

TABLE 5.5: Climatic variables used in the analysis of the HGDP data set.

BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max - min)
BIO3	Isothermality (BIO2/BIO7)
BIO4	Temperature Seasonality (standard deviation * 100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter

A total of 2,624 (0.4%) SNPs obtained $|z|$-scores greater than 5 (Supplementary Tables 5.6 and 5.7). The cutoff $|z| > 5$ ($P < 10^{-7}$) corresponds to the standard Bonferroni correction for a nominal value of type I error $\alpha < 0.01$ and L of order 10^5. Among loci with $|z|$-scores greater than 5, 28 genome-wide association study (GWAS) SNPs with known disease or trait association were found (Hindorff et al., 2009). These include several SNPs discovered by Hancock et al. (2011). For example, the SNPs rs12913832 and rs28777 have $|z|$-scores greater than 6 and are associated with genes OCA2 and SLC45A2 (table 5.8). Among the SNPs significantly correlated with climatic gradients, several notable examples include genes associated with celiac disease (ICOSLG), height (LHX3-QSOX2 and IGF1), and vitamin D synthesis or activation (NADSYN1-encoding nicotinamide adenine dinucleotide synthetase and DHCR7 the gene encoding 7-dehydrocholesterol reductase, an enzyme catalyzing the production in skin of cholesterol from 7-dehydrocholesterol) (table 5.8).

TABLE 5.6: Human data. HGDP SNPs with z-scores with absolute value greater than 7 in genes with molecular (Mol), and biological (Bio) functions associated with these genes.

SNP	CHR	BP	Gene	—z—-score	Gene function
7529482	1	203659355	ATP2B4/intron	7.15	(Mol) calmodulin binding ; protein binding ; hydrolase activity ; calcium-transporting ATPase activity ; metal ion binding ; nucleotide binding ; ATP binding ; hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances ; PDZ domain binding (Bio) platelet activation ; transport ; ATP biosynthetic process ; blood coagulation ; transmembrane transport ; cation transport
3816186	2	42936547	MTA3/nearGene-3	7.35	(Mol) zinc ion binding ; sequence-specific DNA binding ; metal ion binding ; transcription factor activity
4681618	3	150146026	TSC22D2/intron	7.25	(Mol) transcription factor activity
9784335	3	150159767	TSC22D2/intron	7.21	(Bio) response to osmotic stress
10935800	3	150149696	TSC22D2/intron	7.36	
11708779	3	55934939	ERC2/intron	7.40	(Mol) protein binding
144173	7	100416250	EPHB4/cds-synon	8.90	(Mol) protein binding ; protein-tyrosine kinase activity ; ephrin receptor activity ; nucleotide binding ; transmembrane receptor protein tyrosine kinase activity ; receptor activity ; ATP binding heart morphogenesis ; cell migration during sprouting angiogenesis ; protein amino acid autophosphorylation ; multicellular organismal development ; ephrin receptor signaling pathway ; angiogenesis ; cell adhesion
3807496	7	16821355	TSPAN13/intron	7.46	–
4729616	7	100462565	SLC12A9/intron	7.35	(Mol) cation :chloride symporter activity (Bio) transmembrane transport
6942733	7	100350763	ZAN/missense	7.41	(Bio) cell-cell adhesion ; binding of sperm to zona pellucida
10953303	7	100365613	ZAN/missense	7.37	
989465	8	32105334	NRG1/intron	7.04	(Mol) protein binding ; transmembrane receptor protein tyrosine kinase activator activity ; growth factor activity ; ErbB-3 class receptor binding ; cytokine activity ; transcription cofactor activity ; protein tyrosine kinase activator activity ; receptor tyrosine kinase binding ; receptor binding
10096233	8	32115256	NRG1/intron	7.15	(Bio) nervous system development ; regulation of protein heterodimerization activity ; Notch signaling pathway ; positive regulation of cell adhesion ; transmembrane receptor protein tyrosine kinase activation (dimerization) ; neural crest cell development ; cellular protein complex disassembly ; wound healing ; regulation of protein homodimerization activity ; ventricular cardiac muscle cell differentiation ; positive regulation of striated muscle cell differentiation ; positive regulation of cell growth ; cardiac muscle cell differentiation ; cell proliferation ; embryonic development ; mammary gland development ; anti-apoptosis ; cell communication ; negative regulation of secretion ; negative regulation of transcription, DNA-dependent ; transmembrane receptor protein tyrosine kinase signaling pathway ; positive regulation of cardiac muscle cell proliferation
10756461	9	13185149	MPDZ/intron	7.79	(Mol) protein C-terminus binding ; protein binding (Bio) interspecies interaction between organisms
1538677	10	72543579	C10orf27/intron	8.11	(Bio) multicellular organismal development ; spermatogenesis ; cell differentiation
12415051	10	72543913	C10orf27/intron	8.36	
10998340	10	70383593	TET1/intron	8.24	(Mol) oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen ; structure-specific DNA binding ; zinc ion binding ; iron ion binding ; metal ion binding ; oxidoreductase activity (Bio) inner cell mass cell differentiation ; regulation of transcription, DNA-dependent ; stem cell maintenance ; chromatin modification

TABLE 5.7: (continued)

SNP	CHR	BP	Gene	—z—-score	Gene function
2403221	11	9852475	SBF2/intron	8.33	(Mol) protein binding ; protein homodimerization activity ; phosphatase regulator activity ; phosphoinositide binding ; phosphatase binding (Bio) myelination ; protein tetramerization
4910295	11	11311743	GALNTL4/intron	7.27	(Mol) transferase activity, transferring glycosyl groups ; polypeptide N-acetylgalactosaminyltransferase activity ; sugar binding
11066776	12	114264827	RBM19/intron	7.04	(Mol) RNA binding ; nucleotide binding (Bio) positive regulation of embryonic development ; multicellular organismal development
9543476	13	74425228	KLF12/intron	7.16	(Mol) DNA binding ; zinc ion binding ; metal ion binding ; transcription corepressor activity ; transcription factor activity (Bio) regulation of transcription from RNA polymerase II promoter ; positive regulation of transcription from RNA polymerase II promoter ; negative regulation of transcription from RNA polymerase II promoter
1760907	14	20844859	TEP1/intron	7.15	(Mol) telomerase activity ; RNA binding ; nucleotide binding ; ATP binding (Bio) telomere maintenance via recombination
6063071	20	45737763	EYA2/intron	7.09	(Mol) protein binding ; hydrolase activity ; magnesium ion binding ; protein tyrosine phosphatase activity (Bio) istone dephosphorylation ; striated muscle development ; regulation of transcription, DNA-dependent ; apoptosis ; multicellular organismal development ; mesodermal cell fate specification ; chromatin modification ; DNA repair
2294352	22	40827319	MKL1/intron	7.96	(Mol) actin monomer binding ; leucine zipper domain binding ; protein binding ; nucleic acid binding ; transcription coactivator activity
3827382	22	40881403	MKL1/intron	7.73	(Bio) positive regulation of transcription, DNA-dependent ; anti-apoptosis ; smooth muscle cell differentiation ; positive regulation of transcription from RNA polymerase II promoter
6001912	22	40828361	MKL1/intron	7.26	
6001913	22	40836753	MKL1/intron	7.51	
17002034	22	40996367	MKL1/intron	8.01	
5917471	X	37652518	CYBB/intron	7.04	(Mol) protein binding ; FAD binding ; electron carrier activity ; protein heterodimerization activity ; metal ion binding ; superoxide-generating NADPH oxidase activity ; heme binding ; voltage-gated ion channel activity ; oxidoreductase activity (Bio) respiratory burst ; superoxide metabolic process ; innate immune response ; ion transport ; inflammatory response ; superoxide release ; hydrogen peroxide biosynthetic process
5944708	X	25000842	POLA1/intron	7.63	(Mol) DNA primase activity ; metal ion binding ; nucleotide binding ; DNA-directed DNA polymerase activity ; nucleotidyltransferase activity ; transferase activity ; protein binding ; DNA binding ; protein heterodimerization activity ; purine nucleotide binding ; double-stranded DNA binding ; nucleoside binding ; chromatin binding ; pyrimidine nucleotide binding (Bio) DNA replication initiation ; M/G1 transition of mitotic cell cycle ; interspecies interaction between organisms ; DNA replication, synthesis of RNA primer ; DNA strand elongation during DNA replication ; leading strand elongation ; DNA repair ; lagging strand elongation ; double-strand break repair via nonhomologous end joining ; telomere maintenance via semi-conservative replication ; G1/S-specific transcription in mitotic cell cycle ; cell proliferation ; DNA synthesis during DNA repair ; telomere maintenance via recombination ; nucleobase, nucleoside, nucleotide and nucleic acid metabolic process ; mitotic cell cycle ; DNA replication checkpoint ; S phase of mitotic cell cycle ; DNA replication ; telomere maintenance ; G1/S transition of mitotic cell cycle
6643647	X	153086372	PDZD4/intron	7.95	–

TABLE 5.8: Human data. HGDP SNPs with the highest z-scores among those associated with phenotypic traits in GWAS.

Landscape-Trait Category	Ref. SNP ID	Nearby Gene	Disease or Trait Association	$-\text{Log}_{10}$ (P Value)
Pigmentation and tanning	rs32579	*PPARGC1*	Tanning	9.42
	rs12913832	*OCA2/HERC2*	Eye color, eye color traits, hair color, black vs. blond hair color, black vs. red hair color	9.15
	rs11234027	*DHCR7*	Vitamin D levels	7.78
	rs3129882	*HLA-DRA*	Parkinson's disease	6.97
	rs28777	*SLC45A2*	Black vs. blond hair color, black vs. red hair color	6.90
Immune and autoimmune	rs1250550	*ZMIZ1*	Crohn's disease and inflammatory bowel disease (early onset)	8.77
	rs2735839	*KLK3*	Prostate cancer	8.16
	rs9264942	*RPL3P2*	HIV-1 control	8.02
	rs2179367	Intergenic between *SUMO4* and *ZC3H12D*	Dupuytren's disease	7.57
	rs1551398	Intergenic between *TRIB1* and *LOC100130231*	Crohn's disease	7.45
	rs2289700	*CTSH*	Bipolar disorder	6.98
	rs4819388	*ICOSLG*	Celiac disease	6.67
	rs703842	*CYP27B1/METTL1*	Multiple sclerosis	6.59
	rs12593813	*MAP2K5*	Restless legs syndrome	6.40
	rs4664308	*PLA2R1*	Nephropathy (idiopathic membranous)	6.28
Metabolism	rs10908907	Intergenic *MUC7*	Alcoholism (heaviness of drinking)	8.91
	rs1566039	Intergenic between *PAPD7* and *MIR4278*	Sphingolipid levels	6.89
	rs7665090	*MANBA*	Primary biliary cirrhosis	6.48
Cardiovascular	rs869244	*ADRA2A*	Platelet aggregation	7.20
	rs12034383	*CR1*	Erythrocyte sedimentation rate	7.15
	rs3129882	*HLA-DRA*	Systemic sclerosis	6.97
	rs11897119	*MEIS1*	PR interval	6.71
Height	rs7678436	*NCAPG-LCORL*	Height	9.43
Other	rs12479254	*BOK*	Brain structure	9.43

NOTE.—HGDP SNPs with the highest $|z|$-scores among those associated with phenotypic traits in GWAS.

We performed a Gene Ontology enrichment analysis on human genes with $|z|$-scores greater than 5 (2,624 SNPs). Using a threshold of 0.001 for the false discovery rate (FDR) q-values, we found significant enrichment of gene ontology terms associated with six biological processes linked to cell adhesion and locomotion, neural and organismal development (Supplementary Tables 5.6, 5.7, 5.9). The FDR q-values for the regulation of developmental processes (76 genes) and the regulation of multicellular organismal processes (88 genes) were equal to $q = 0.006$ and $q = 0.003$. For examples of interesting genes with a high level of association with climatic variables, we focus of the 65 SNPs with $|z|$-scores greater than 7. Among the 65 SNPs, EPHB4 ($|z| = 8.90$) is involved in heart morphogenesis and angiogenesis, NRG1 ($|z| = 7.15$) is involved with nervous system development and cell proliferation, RBM19 ($|z| = 7.04$) is involved with positive regulation of embryonic development, EYA2 ($|z| = 7.09$) is involved with eye development and DNA repair, and POLA1 ($|z| = 7.63$) is involved with the mitotic cell cycle and cell proliferation (Saccone et al., 2011; Hornbeck et al., 2012) (Supplementary Table 5.6). A supplementary table, available online, describes a list of 508 SNPs with $|z|$-scores greater than 6.

GO Term	Description	FDR q-value	Enrichment	No. genes
GO :0050793	regulation of developmental process	3.07e-3	1.75	76
GO :0007411	axon guidance	4.3e-3	2.7	30
GO :0007155	cell adhesion	4.46e-3	2.02	49
GO :0040011	locomotion	4.46e-3	1.97	59
GO :0022610	biological adhesion	5.57e-3	2.02	49
GO :0051239	regulation of multicellular organismal process	6.59e-3	1.61	88

TABLE 5.9: Significant enrichment of gene ontology terms associated with biological processes.

5.4 Discussion

5.4.1 Interpretation of LFMM results and other methods.

On the basis of a matrix factorization approach, LFMMs provide a unified framework for estimating effects of environmental and demographic factors on genetic variation. Without environmental variables, LFMMs are similar to performing a sparse version of a probabilistic PCA of allele frequencies (Tipping and Bishop, 1999; Engelhardt and Stephens, 2010). When environmental variables are included, hidden factors capture the part of genetic variation that cannot be explained by the set of measured environmental variables. This fraction of genetic variation could result from the demographic history of the species, unknown environmental pressures or from IBD patterns.

Although a plethora of statistical tests have been proposed for detecting genes evolving under positive selection and local adaptation (Storz, 2005; Novembre and Di Rienzo, 2009), the development of tests based on correlations with habitat or landscape variables is still recent (Joost et al., 2007; Hancock et al., 2008). Compared with methods based on summary statistics, tests based on environmental association have increased power to detect selection from standing genetic variation and soft sweeps in a species genome (Pritchard et al., 2010; Schoville et al., 2012). However, simple implementation of these tests, for example, linear or logistic regression models, can be misleading in the presence of IBD patterns (Meirmans, 2012). Our simulation results provide clear evidence that tests based on LFMMs significantly reduce the rates of FP associations in the presence of IBD.

Rates of FP and FN were also investigated for three regression methods that include corrections for population genetic structure : PMTs, PCRM, and standard linear mixed models. In the case of phylogenetic comparative analyses that infer environmental correlations for correlated DNA sequences, PMTs were reported to be erroneous (Harmon and Glor, 2010). In addition, Legendre and Legendre (2012) warn against using partial Mantel correlations. The high error rate may stem from autocorrelation of matrix elements due to underlying phylogenetic structure. We found that PMTs produced an excess of high

and low P-values under IBD assumptions. Although PMTs were not correctly calibrated, these tests can provide a useful statistic for ranking loci, and they can detect interesting associations after choosing a tail cutoff (Fumagalli et al., 2011). P-values based on PCRM and standard linear mixed models were correctly calibrated. But we found that the three regression methods had low power to detect true associations under IBD assumptions. Although these approaches might be useful to detect alleles with strong associations to environmental gradients, they can miss several interesting associations. FN rates were high for PMT, PCRM, and `GEMMA` because the simulated environmental variable was strongly correlated with population structure. We suspect all regression methods -including LFMM- have higher power when environmental gradients are uncorrelated to the main axis of neutral genetic variation.

Both the mixed model approach of the computer program `BAYENV` and the LFMM approach include a covariance structure in a regression model, but there are important differences between the two approaches. A first improvement is that LFMMs estimate latent factors and regression coefficients simultaneously, whereas `BAYENV` first estimates a covariance matrix, and then uses it when estimating (random) environmental effects. To apply `BAYENV`, the authors suggest utilizing selectively neutral SNPs to estimate the covariance matrix. Inclusion of adaptive markers in the "neutral set" is sometimes unavoidable, and in this case, methods based on the empirical covariance matrix may overlook interesting associations. For Loblolly pines expressed sequence data, the distinction between the two approaches may explain the differences we observed between the list of loci obtained from LFMM and the list obtained from `BAYENV` (Eckert et al., 2010). For the pine data, it was difficult to select neutral SNPs from the background a priori. Another distinction between LFMM and `BAYENV` approaches is our use of low rank approximations of the covariance matrix. LFMMs actually estimate correlations between environmental predictors and allele frequencies while K hidden factors explain residual genetic variation, where K is much smaller than the sample size. Though program speed is generally difficult to evaluate for Markov chain Monte Carlo methods, we observed that LFMM was computationally faster than `BAYENV` when analyzing large data sets.

5.4.2 Number of Latent Factors

A potential weakness of tests based on LFMM is that we need to choose K. In the LFMM modeling approach, the choice of low values for K is important for optimizing the computational performances of the estimation algorithm. This choice is reminiscent of selecting the number of components in PCA or in Bayesian clustering programs, and it has also an impact on test outcomes. For values of K taken too large, the tests are conservative,

and the power to reject neutrality declines. Estimates of K that minimize the trade-off between the bias and variance for our statistical estimates could be obtained by using cross-validation procedures. Cross-validation procedures are computationally intensive, so instead we use Tracy–Widom theory to select K (Patterson et al., 2006). We evaluated this choice during our simulation analysis and found that P-values were well calibrated. Although the choice of Tracy–Widom estimates is suboptimal, the performances of LFMMs were superior to those of `BAYENV` in simulations of IBD patterns. In the analysis of human data, we restricted K to be less than 50 (approximately the number of population samples). We suggest that, when there is a reasonable estimate of the number of genetic clusters for a species, this should be used in LFMM tests directly. For example, estimates of K based on independent genetic data sets could be obtained from Bayesian clustering programs like `STRUCTURE` (Pritchard et al., 2000a). Although finer grain population structure could also be evaluated (Lawson et al., 2012), our choice was again motivated by a trade-off between accuracy and run-time. A future development of our LFMM approach will be to develop fast numerical optimization procedures based on variational approximations of the likelihood, which will allow us to implement cross-validation algorithms and increase the power of tests.

5.4.3 Plant and Human Data

For *Pinus taeda*, the LFMM results confirmed that several ESTs previously discovered with `BAYENV` had functions linked to climate (Eckert et al., 2010). In addition, the LFMM analysis discovered new interesting candidate SNPs. Those variants include functions associated with wound repair and immunity ; photosynthetic activity and carotenoid biosynthesis ; cellular respiration and carbohydrate metabolism ; and heat, salt, and oxidative stress responses (table 5.3). Applying LFMMs to the HGDP data, we found that a total of 0.4% of all polymorphisms (2,624 SNPs) exhibited significant associations with temperature gradients ($|z| > 5$). For example, we identified SNPs associated with the gene OCA2 that may be functionally linked to blue or brown eye color and the gene SLC45A2 that may be associated with skin pigmentation (Hancock et al., 2011). This list also contained SNPs identified from GWAS studies of height and vitamin D synthesis and diseases such as gluten intolerance and Crohn's disease. Another interesting result is that the list of genic SNPs with $|z|$-scores greater than 5 ($|z| > 5$) was enriched for gene ontology terms associated with six biological processes linked to cell adhesion and locomotion, neural and organismal development. Among the highest scores, the genes EPHB4, BOK, and NRG1 −with functions related to heart and brain development− were associated with

climatic gradients. Although cautious interpretations of the results may be required (Pavlidis et al., 2012), our data analysis confirmed that many allele frequencies correlate with climatic gradients or with some evolutionary pressures associated with these gradients.

5.4.4 Conclusion

With ever increasing amounts of genetic data generated by high-throughput sequencing technologies, population genetic methods have shifted from empirical approaches to models that incorporate hidden factors. Estimates of ancestry and other population parameters are commonly obtained from mixture models (Pritchard et al., 2000a; Durand et al., 2009; Alexander and Lange, 2011), principal component analyses (Patterson et al., 2006), hidden Markov models (Price et al., 2009), and factor analysis (Engelhardt and Stephens, 2010). Our study contributes to the factor analysis methods for population and landscape genomic analysis by implementing new tests of gene-environment association. These new tests use comparisons between closely related populations that have adapted to different environments, and they may help to detect modes of selection that differ from the classic selective sweep paradigm.

5.5 Materials & Methods

5.5.1 LFMM Implementation Details.

Consider the data matrix, $(G_{i\ell})$, where each entry records the allele frequency in individual i at the genomic locus ℓ, $1 \leq i \leq n$, $1 \leq \ell \leq L$, and n and L represent the total sample size and number of loci, respectively. LFMMs were defined by the following equation :

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell}$$

where μ_ℓ is a locus-specific effect, β_ℓ is a d-dimensional vector of regression coefficients, U_i and V_ℓ are scalar vectors with K dimensions ($1 \leq K \leq n$). The residuals $\epsilon_{i\ell}$ are statistically independent Gaussian variables of mean zero and variance σ^2.

We use Bayesian analysis to estimate the regression coefficients and their standard deviations. We assume Gaussian prior distributions on μ_ℓ and $\beta_{\ell j}$ with means equal to zero and variances σ_μ^2 and $\sigma_{\beta_j}^2$ ($\beta_{\ell j} \sim \mathrm{N}(0, \sigma_{\beta_j}^2)$). Prior distributions on U_i and V_ℓ are Gaussian distributions with means equal to zero and constant variance for each component (the

components are independent random variables). The variance of V_ℓ is set to $\sigma_V^2 = 1$. The prior distributions on σ_μ^2 and $\sigma_{\beta_j}^2$ are noninformative distributions. The variance of each factor, σ_U^2, follows an inverse-Gamma distribution $\Gamma^{-1}(\eta, \eta)$ where $\eta = 10^2 - 10^3$. This parameterization encourages sparsity in factor estimates and provides a more accurate description of underlying population structure (Engelhardt and Stephens, 2010).

To simultaneously estimate scores (U) and loadings (V), environmmental effects (β), and biases (μ), we implemented a Gibbs sampler algorithm for LFMMs (Supplementary File S1, section 5.6). The Gibbs sampler was based on computing products of matrices of low dimension –typical values of K were less than 50 –and its speed scales with the current size of SNP data sets, around $n \approx 1,000$ and $L \approx 500,000$. We implemented a stochastic algorithm to compute standard deviations for the environmental effects (Supplementary File S1, section 5.6). The $|z|$-scores were computed as the ratios between the centered values of the regression coefficients β_ℓ and their standard deviations, and they were converted into P-values according to the standard Gaussian distribution. The cutoff for $|z|$-scores was obtained after applying a Bonferroni correction, corresponding to a type I error of 0.01. From a preliminary set of experiments using data simulated from the model defined in equation (5.1), we found that the estimates of fixed effects stabilized quickly, after 1,000 to 10,000 sweeps for $n = 100 - 1,000$ individuals and $L = 1,000 - 100,000$ loci. A 10-fold increase in the number of sweeps, however, was necessary to recover the true values of the latent factors. Additionally, we developed numerical optimization methods to compute *maximum a posteriori* (MAP) estimates for the LFMM. One of these methods, the alternate least square method uses deterministic steps that are similar to our stochastic Gibbs sampler (Koren et al., 2009). When checking for convergence of the Markov chain Monte Carlo (MCMC) algorithm, we also found that least square estimates of regression coefficients were close to the point estimates computed by the Gibbs sampler method. The computational complexity of a single sweep of the LFMM Gibbs sampler algorithm is of order $O(nLK^3)$. For about 1,000 loci and 1,000 individuals, the LFMM MCMC algorithm was run for approximately 1 minute of a 2.4 GHz Intel Xeon 64 bit processor. For larger data sets with 650 K loci and 1,000 individuals, we used a multithreated version of the algorithm, for which a single run lasted approximately 24 h on a multiprocessor computer system (using 10 threads).

5.5.2 Alternative Regression Approaches

The standard linear regression model (LM) was defined as

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \epsilon_{i\ell}\,. \tag{5.7}$$

and the GLMs used the binomial family and the canonical link. The PCRM was defined as

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \tilde{U}_i^T V_\ell + \epsilon_{i\ell}\,, \tag{5.8}$$

where $(\tilde{U}_i)$ are the first K PCs computed from the matrix G. For each SNP, we applied PMTs to assess the relationship between the matrix of allele frequency distances and the matrix for environmental variable distance (Smouse et al., 1986; Legendre and Legendre, 2012). PMTs are nonparametric permutation-based tests for quantifying association between two distance matrices, while controlling for the effect of a third matrix. The allele frequency distance matrices were computed using Nei's distance (Nei, 1972), and the environmental distance matrix used the Euclidean distance. The third matrix was the Pearson's correlation matrix computed over all loci. P-values were computed from the R package `vegan` using 10,000 permutations (R Development Core Team., 2012). We used the computer program `GEMMA` to implement a standard linear mixed model for genome-wide association studies (Zhou and Stephens, 2012). The model had the following form

$$X_i = \sum_\ell G_{i\ell}\beta_\ell + u_i + \epsilon_i \quad i = 1, \ldots, n\,,$$

where u is a multivariate random effect having a Gaussian distribution of covariance matrix $\lambda\tau^{-1}\Lambda$, and each ϵ_i is a residual error vector having a Gaussian distribution of variance τ^{-1}. The parameter τ is the variance of the residual errors, and λ is the ratio between the variance components. The matrix Λ is an $n \times n$ relatedness matrix. Finally, we used the generalized linear mixed model implemented in the computer program `BAYENV` with the default settings of the software (Coop et al., 2010).

5.5.3 LFMM Generative Model Simulations

We used equation (5.1) with $\beta = 0$ to generate data under a null hypothesis of no association with any environmental variables. In these experiments, we set the number of individuals to $n = 100$, and the number of loci to $L = 1,000$. We used six values, $K = 1, 3, 5, 7, 10$ and 20, for the rank of the factor matrix, V. For each series of experiments, we generated 10 replicates of this generative model, and we studied the distributions of

P-values for tests using LFMMs. In these tests, we set the rank of the factor matrix equal to the values we used to generate simulations.

Next we used equation (5.1) to generate data showing various levels of population structure and association with an environmental variable. The environmental variable was uniformly generated in the range $(0, 1)$. Here, we used three values for the rank of the factor matrix, $K = 2, 20$ and 100, representing low, moderate, and high levels of underlying population genetic structure. For each series of experiments, we generated 20 replicates of the generative model.

To compute point estimates of environmental effects and their $|z|$-scores, Gibbs sampler algorithms were run for 1,000 sweeps after a burn-in period of 100 sweeps. For these particular run length parameters, we checked that similar estimates were obtained for distinct initializations of the algorithm. For each locus, we recorded both the true, β_ℓ, and estimated environmental effects, $\hat{\beta}_\ell$, and evaluated the absolute error Formula

$$E_\ell = \left| \beta_\ell - \hat{\beta}_\ell \right| .$$

5.5.4 Spatial Coalescent Simulations

To enable comparisons with other models, we simulated genotypic data from spatial coalescent models with the computer program `ms` (Hudson, 2002). Ten data sets were generated according to a linear stepping-stone model with 40 demes, setting the effective migration rate between pairs of adjacent demes to the value $4Nm = 25$. Sampling five individuals in each deme, each data set included a total of $n = 200$ haploid individuals genotyped at $L = 1,000$ unlinked SNP loci. We ran the LFMM during 100 sweeps for burn-in, and we used the next 900 sweeps to compute point estimates, variances, and $|z|$-scores. An environmental variable, x, was defined for each population as the geographic identifier of the population in the linear stepping-stone model.

We created an environmental gradient for the artificial variable x using a logistic function, $s(x)$, of x as follows

$$s(x) = \frac{1}{1 + e^{\theta(x-20)}} , \quad \theta > 0. \tag{5.9}$$

For each of the 10 previously generated neutral stepping-stone simulations, we simulated binary alleles at 50 unlinked loci for each deme x with frequency $s(x)$, and with the slope of the gradient $\theta = 0.1 - 0.2$. We then obtained 10 data sets with $L = 1050$ unlinked loci

including 50 loci correlated with the environmental gradient, $s(x)$. Using Tracy–Widom tests implemented in SmartPCA, we found that the number of principal components with P-values smaller than 0.01 was around $K_{\text{TW}} = 7$. Using the Bayesian clustering programs `STRUCTURE` and `TESS`, we found that $K = 5$ components could better describe our simulated data. A value $\theta = 0.2$ corresponds to a strong intensity of selection through geographic space, whereas $\theta = 0.1$ corresponds to a weak intensity of selection. We used the value $\theta = 0.2$ when comparing tests based on linear and PC regression models. When comparing LFMMs with `BAYENV`, we used the value $\theta = 0.1$ to better fit the objectives of both models. As `BAYENV` returns Bayes factors instead of P-values, we considered ranked lists recording the M loci corresponding to the strongest associations (M between 1 and $L = 1,050$). For each M, we computed the number of TPs and the number of FPs. Locus ranking was performed on the basis of $|z|$-scores in LFMM and on the basis of Bayes factors in `BAYENV`. The LFMM tests used values of K equal to $K = 1, 3, 5, 7, 10$ and 20, and we used of the `BAYENV` algorithm to compute Bayes factors. Experiments were assessed by counting the number of FP and FN associations, and by measuring the AUC averaged over 10 replicates.

5.5.5 Real Data

5.5.5.1 Loblolly Pine

The Loblolly pine data consisted of 1,730 SNPs selected in ESTs for 682 individuals (Eckert et al., 2010). We considered 5 environmental variables representing the 5 first components of a PCA for 60 climatic variables (Eckert et al., 2010). The first component (PC1) was mainly described by latitude, longitude, temperature, and winter aridity. PC2 was described by longitude, spring-fall aridity, and precipitation (Eckert et al., 2010). For each of the 5 environmental variables, we applied the LFMM algorithm using 100 sweeps for burn-in and 400 additional sweeps to compute $|z|$-scores for all loci. On the basis of a prior analysis of the genotypic data with the program `smartPCA` and Tracy–Widom tests, we used $K = 10$ latent factors.

5.5.5.2 Human Data

Genotypes from the HGDP–CEPH data set were generated on Illumina 650 K arrays (Li et al., 2008), and the data were filtered to remove low quality SNPs included in the original files. We extracted climatic data for each of the 52 population samples using the WorldClim data set at 30 arcsecond (1km^2) resolution (Hijmans et al., 2005). These data

included 11 bioclimatic variables interpolated from global weather station data collected during a 50-year period (averaged of the years 1950-2000. The environmental variables were mainly related to temperature data. These variables included annual mean temperature, mean diurnal range, maximum temperature of warmest month, minimum temperature of coldest month, and so forth (Supplementary Table 5.5). We summarized them by using the first axis of a PCA (all 11 climatic variables were given similar loadings). For this environmental proxy, we applied the LFMM algorithm and computed $|z|$-scores for each locus, using 100 sweeps for burn-in and 900 additional sweeps to compute estimates. We used $K = 50$ which was of the same order as the number of population samples and the value returned by the Tracy–Widom tests. We investigated whether the gene ontology terms of environmentally associated SNPs were enriched in specific categories of biological processes. The list of target genes with $|z|$-scores greater than 5 was compared with the background list of 14,042 genes represented in the HGDP–CEPH data set using a hypergeometric distribution. This test was implemented using the `GORILLA` software tool (Eden et al., 2009), with significance determined by at an FDR corrected q-value threshold of 0.01.

5.5.6 Software availability

Source codes and computer programs for fitting LFMMs are available from the author websites (http ://membres-timc.imag.fr/Eric.Frichot/ and
http ://membres-timc.imag.fr/Olivier.Francois/lfmm.html).

5.5.7 Acknowlegments.

This work was supported by la Région Rhône-Alpes grant to E.F. and O.F. ; National Science Foundation grant OISE-0965038 to S.D.S. ; and Grenoble INP grant to O.F. The authors are grateful to Florian Alberto, Pierre De Villemereuil, and Oscar Gaggiotti for useful comments of the LFMM software and to Daniel Wegmann and Matteo Fumagalli for their careful reading and helpful suggestions on a previous version of the manuscript.

5.6 Supporting text : Gibbs Sampling algorithm for latent factor mixed models

Let us denote by n the total sample size, L is the number of loci, and D is the dimension of the set of environmental covariates. We denote $\mathrm{N}(\mu, \Sigma)$ the multivariate Gaussian distribution of mean μ and of covariance matrix Σ, and $\Gamma^{-1}(a, b)$ is the inverse-gamma distribution of shape a and rate b (and scale $1/b$).

Prior distributions. The prior distributions on the LFMM parameters are defined hierarchically as follows. For all i, ℓ, the allele count is described by

$$G_{i,\ell} \mid U_i, V_\ell, \beta_\ell, \mu_\ell, \sigma^2 \sim \mathrm{N}(X_i\beta_\ell + \mu_\ell + U_i^T V_\ell, \sigma^2)\,, \tag{5.10}$$

where μ_ℓ is a locus-specific mean, and σ^2 is a residual variance term. The factors are described as

$$U_i|\sigma_U^2 \sim \mathrm{N}(0, \sigma_U^2 \mathrm{I}_K)\,, \tag{5.11}$$

where I_K is the identity matrix with K dimensions, and K is the number of factors in the model. The scores are described as

$$V_\ell \sim \mathrm{N}(0, \mathrm{I}_K)\,, \tag{5.12}$$

where I_K is the identity matrix with K dimensions. For $1 \leq j \leq D$, the fixed effect regression coefficients are described by

$$\beta_{j\ell}|\sigma_{\beta_j}^2 \sim \mathrm{N}(0, \sigma_{\beta_j}^2)\,, \tag{5.13}$$

and the intercept coefficients are described by

$$\mu_\ell|\sigma_\mu^2 \sim \mathrm{N}(0, \sigma_\mu^2)\,. \tag{5.14}$$

The hyper-parameters, $\sigma_{\beta_j}^2$, σ_μ^2 follow non-informative inverse-chi squared distributions, and σ_U^2 follows an inverse-gamma distribution $\Gamma^{-1}(\eta, \eta)$ where $\eta = 10^2 - 10^3$. This parameterization encourages sparsity in factor estimates.

Conditional distributions. The LFM model is a hierarchical model with conditional distributions that can be described as follows :

$$p(\sigma_U^2|U,\eta) = \Gamma^{-1}(\eta + \frac{nK}{2}, \frac{1}{2}\sum_i U_i^T U_i + \eta) \tag{5.15}$$

$$p(\sigma_{\beta_j}^2|\beta) = \Gamma^{-1}(1 + \frac{L}{2}, \frac{1}{2}\sum_l \beta_{jl}^2 + 1) \tag{5.16}$$

$$p(\sigma_\mu^2|\mu) = \Gamma^{-1}(1 + \frac{L}{2}, \frac{1}{2}\sum_\ell \mu_\ell^2 + 1) \tag{5.17}$$

$$p(U_i|G,V,\beta,\mu,\sigma_U^2,\sigma^2) = \mathrm{N}(\mu_U^i, {\Delta_U^i}^{-1})\,, \tag{5.18}$$

where

$$\Delta_U^i = {\sigma_U}^{-2}\mathrm{I}_K + \sigma^{-2}\sum_\ell V_\ell V_\ell^T\,, \tag{5.19}$$

and

$$\mu_U^i = \sigma^{-2}(\Delta_U^i)^{-1}\sum_\ell (G_{i,\ell} - X_i\beta_\ell - \mu_\ell)V_\ell\,. \tag{5.20}$$

In addition, we have

$$p(V_\ell|G,U,\beta,\mu,\alpha_G) = \mathrm{N}(\mu_V^\ell, {\Delta_V^\ell}^{-1})\,, \tag{5.21}$$

where

$$\Delta_V^\ell = {\sigma_V}^{-2}\mathrm{I}_K + \sigma^{-2}\sum_i U_i U_i^T \tag{5.22}$$

and

$$\mu_V^\ell = \sigma^{-2}(\Delta_V^\ell)^{-1}\sum_i (G_{i,\ell} - X_i\beta_\ell - \mu_\ell)U_i\,. \tag{5.23}$$

We have

$$p(\beta_\ell|G,U,V,\mu,\sigma_\beta(1)^2,\ldots,\sigma_\beta(d)^2,\sigma^2) = \mathrm{N}(\mu_\beta^\ell, {\Delta_\beta^\ell}^{-1}) \tag{5.24}$$

where

$$\Delta_\beta^\ell = \mathrm{diag}({\sigma_{\beta_1}}^{-2},\cdots,{\sigma_{\beta_j}}^{-2}) + \sigma^{-2}\sum_i X_i^T X_i \tag{5.25}$$

and

$$\mu_\beta^\ell = \sigma^{-2}(\Delta_\beta^\ell)^{-1}\sum_i (G_{i,\ell} - U_i^T V_\ell - \mu_\ell)X_i^T\,. \tag{5.26}$$

Finally, we have

$$p(\mu_\ell|G,U,V,\beta,\sigma_\mu^2,\sigma^2) = \mathrm{N}(\mu_\mu^\ell, {\Delta_\mu^\ell}^{-1}) \tag{5.27}$$

where

$$\Delta_\mu^\ell = {\sigma_\mu}^{-2} + n\sigma^{-2} \tag{5.28}$$

and

$$\mu_{\mu}^{\ell} = \sigma^{-2}(\Delta_{\mu}^{\ell})^{-1} \sum_{i} (G_{i,\ell} - U_i^T V_\ell - X_i \beta_\ell) \tag{5.29}$$

The parameter σ^2 is updated at each iteration using the current residual variance. The other parameters are updated through Gibbs sampling cycles.

Main algorithm Let nc be the number of Gibbs sampler cycles, and 'burn' be the number of cycles used for burn-in.

1. Initialize the model parameters

$$U = 0_{K,n}$$

$$V = 0_{K,L}$$

$$\beta = 0_{L,D}$$

$$\mu = 0_{L,1}$$

2. For $t = 1, \ldots,$ nc
   - Input missing values at locus ℓ for individual i,

$$G_{i,\ell} \leftarrow {U_i}^{(t-1)^T} V_\ell^{(t-1)} + X_i^{(t-1)} \beta_\ell^{(t-1)}$$

   - Update the residual variance

$$\sigma^{2(t)} = var(G - {U^{(t-1)}}^T V^{(t-1)} - X^{(t-1)} \beta^{(t-1)})$$

   - Sample the hyper-parameters

$${\sigma_U^2}^{(t)} \sim p(\sigma_U^2 | U^{(t-1)}, \eta)$$

$${\sigma_\beta^2}^{(t)} \sim p(\sigma_\beta^2 | \beta^{(t-1)})$$

$${\sigma_\mu^2}^{(t)} \sim p(\sigma_\mu^2 | \mu^{(t-1)})$$

   - For each locus ℓ, sample

$$\mu_\ell^{(t)} \sim p(\mu_\ell | G, U^{(t-1)}, V^{(t-1)}, \beta^{(t-1)}, {\sigma_\mu^2}^{(t)}, \sigma^{2(t)})$$

$$\beta_\ell^{(t)} \sim p(\beta_\ell | G, U^{(t-1)}, V^{(t-1)}, \mu^{(t)}, {\sigma_{\beta_1}^2}^{(t)}, \ldots, {\sigma_{\beta_j}^2}^{(t)}, \sigma^{2(t)})$$

   - For each individual i, sample

$$U_i^{(t)} \sim p(U_i | G, \mu^{(t)}, V^{(t-1)}, \beta^{(t)}, {\sigma_U^2}^{(t)}, \sigma^{2(t)})$$

   - For each locus ℓ, sample

$$V_\ell^{(t)} \sim p(V_\ell | G, \mu^{(t)}, U^{(t)}, \beta^{(t)}, \sigma^{2(t)})$$

3. compute the parameters

$$U = \text{mean}(U^{(\text{burn}+1)}, \ldots, U^{(\text{nc})})$$

$$V = \text{mean}(V^{(\text{burn}+1)}, \ldots, V^{(\text{nc})})$$

$$\beta = \text{mean}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})})$$

$$\mu = \text{mean}(\mu^{(\text{burn}+1)}, \ldots, \mu^{(\text{nc})})$$

$$Z = \text{mean}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})}) / \text{var}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})})^{\frac{1}{2}}$$

Chapitre 6

Extensions et perspectives statistiques

Durant cette thèse, nous avons proposé, dans un cadre bayésien, des nouvelles approches pour l'étude de la variation génétique d'un ensemble d'individus d'une espèce. En particulier, nous avons proposé de nouveaux modèles de régression corrigeant les facteurs de confusion grâce à l'introduction de facteurs latents.

Pour effectuer des tests d'association écologique, nous avons proposé d'estimer des coefficients de régression du modèle LFMM par un échantillonnage de Gibbs. Cet algorithme bayésien a de nombreux avantages. Il permet d'obtenir une estimation de la loi a posteriori des coefficients de régression et de pouvoir effectuer un test. Toutefois, l'algorithme d'échantillonnage de Gibbs a aussi certaines limites. Tout d'abord, l'algorithme peut être coûteux en calcul. De plus, l'algorithme de Gibbs pour le modèle LFMM demande de faire un certain nombre de choix de modélisation, comme celui des lois a priori (Zhou et al., 2013). Enfin, en grande dimension, l'algorithme de Gibbs parcourt seulement un sous-ensemble de l'espace d'état dépendant de l'initialisation de l'échantillonneur de Gibbs (West, 2003).

L'objectif de ce chapitre d'extensions et de perspectives est de proposer des alternatives aux solutions mises en place pour estimer les coefficients du modèle LFMM. En effet, le modèle LFMM a été construit à partir de la régression linéaire et des modèles à facteurs latents. Ces deux types de modèles ont été largement étudiés dans le passé (Anderson and Gerbing, 1984; McCullagh and Nelder, 1989; Jolliffe, 1986). On connait de nombreux algorithmes d'estimation pour la régression et les modèles à facteurs latents. On s'appuiera donc sur les résultats connus pour ces deux types de modèles pour envisager d'autres

estimateurs des coefficients de régression du modèle LFMM ainsi que les algorithmes d'estimation associés.

Dans un premier temps, nous présentons les différentes approches d'estimation des paramètres du modèle LFMM. Nous présentons des approches fréquentistes, des approches probabilistes et des approches bayésiennes. Puis, pour chaque approche, nous décrivons des résultats préliminaires. Tout d'abord, nous présentons nos résultats pour les approches fréquentistes avec des estimateurs de maximum de vraisemblance et un estimateur de maximum vraisemblance régularisée de type régression "Ridge". Ensuite nous nous intéressons aux résultats pour des approches probabilistes avec un estimateur par algorithme Expectation-Maximisation (EM) et un estimateur par maximisation d'une vraisemblance complète. Enfin, nous présentons nos résultats pour des approches bayésiennes avec un estimateur de maximum a posteriori (MAP) et un estimateur par approximation variationnelle bayésienne (VB). Nous terminons ce chapitre par une discussion sur les avantages et les limites des différents estimateurs proposés et sur les efforts à mettre en œuvre pour les implanter.

6.1 Rappel des notations

On note G la matrice des génotypes de taille $n \times L$ où n est le nombre d'individus et L est le nombre de SNPs. Dans toute cette partie, pour faciliter les notations, on suppose que la matrice G est centrée pour chaque SNP. En plus des données génotypiques, on note X la matrice d'indicateurs environnementaux de taille $n \times d$ où d est le nombre d'indicateurs environnementaux. On suppose que X est de rang d. De plus, on note $vect(X)$ le sous-espace vectoriel engendré par X dans $\mathbb{R}^n$ et $vect(X^{\perp})$ son supplémentaire dans $\mathbb{R}^n$.

On note $||.||_F$ la norme de Frobenius. On note I_n la matrice identité de taille $n \times n$. On note $tr(A)$ la trace de la matrice A. On note $svd_K(G)$ l'approximation de rang K de la matrice G par sa décomposition en valeurs singulières.

On note, P_X, la matrice de projection orthogonale sur $vect(X)$,

$$P_X = X(X^T X)^{-1} X^T.$$

De plus, on note, $P_{X^{\perp}}$, la matrice de projection orthogonale sur $vect(X^{\perp})$,

$$P_{X^{\perp}} = I_n - X(X^T X)^{-1} X^T.$$

Comme $vect(X)$ et $vect(X^\perp)$ sont des sous-espaces vectoriels supplémentaires dans $\mathbb{R}^n$, on peut décomposer de manière unique tout vecteur $u \in \mathbb{R}^n$ comme somme d'un vecteur $u_X = P_X u \in vect(X)$ et d'un vecteur $u_{X^\perp} = P_{X^\perp} u \in vect(X^\perp)$. Par définition d'une projection orthogonale, on ne peut pas calculer u à partir de $u_{X^\perp}$.

Pour contourner cette difficulté, on introduit les matrices de projections obliques P_X^D et $P_{X^\perp}^D$ les projections obliques sur $vect(X)$ et sur $vect(X^\perp)$, pour D une matrice diagonale, de taille $d \times d$ et de diagonale strictement positive

$$P_X^D = X(X^T X + D^{-1})^{-1} X^T, \quad P_{X^\perp}^D = I_n - X(X^T X + D^{-1})^{-1} X^T.$$

De même que précédemment, on peut décomposer de manière unique tout vecteur $u \in \mathbb{R}^n$ comme somme d'un vecteur $u_X^D = P_X^D u \in vect(X)$ et d'un vecteur $u_{X^\perp}^D = P_{X^\perp}^D u \in vect(X^\perp)$. Dans ce cas, comme $P_{X^\perp}^D$ est inversible, on peut déterminer $u = (P_{X^\perp}^D)^{-1} u_{X^\perp}^D$ à partir de $u_{X^\perp}^D$.

6.2 Des approches d'estimation des paramètres du modèle LFMM

Le modèle LFMM s'écrit

$$G_{i\ell} = U_i V_\ell^T + X_i B_\ell^T + \epsilon_{i\ell} \quad i = 1 \dots n, \quad \ell = 1 \dots L,$$

avec $\epsilon_{i\ell}$ un résidu de loi $N(0, \sigma^2)$. La matrice B est la matrice des coefficients de régression de taille $L \times d$. On associe à B un modèle à K facteurs où U la matrice de scores de taille $n \times K$ et V la matrice des poids de taille $L \times K$. On suppose que U et V sont de rang K strictement positif et petit devant n.

Dans cette partie, nous présentons différentes approches afin d'estimer les paramètres du modèle LFMM (Table 6.1). Puis, dans les parties suivantes, nous présenterons les algorithmes et les estimateurs pour ces différentes approches.

6.2.1 Des approches fréquentistes

Nous proposons une estimation des paramètres U, V, B et σ^2 par une approche de maximisation de la vraisemblance et une approche de maximisation de la vraisemblance régularisée.

approche	méthode	variables					
		σ_U^2	D_B	σ^2	U	V	B
		$\Gamma^{-1}(\eta,\eta)$	$\Gamma^{-1}(\eta,\eta)$	$\Gamma^{-1}(\eta,\eta)$	$N(0,\sigma_U^2)$	$N(0,1)$	$N(0,D_B)$
fréquentiste	vraisemblance	–	–	P	P	P	P
	vraisemblance Ridge	–	P	P	P	P	P
probabiliste	EM	–	P	P	P	R	R
	vraisemblance marginale	–	P	P	P	R	R
bayésienne	MAP	P	P	P	R	R	R
	VB	P	P	P	R	R	R
	GS	R	R	R	R	R	R

TABLE 6.1: Synthèse des différents approches. L'abbréviation P signifie que la variable est une paramètre. L'abbréviation R signifie que c'est une variable aléatoire. Un tiret signifie que cette variable n'est pas prise en compte dans la méthode considérée. La variable η est un paramètre de régularisation.

6.2.1.1 Une approche par maximisation de la vraisemblance

L'approche d'estimation des paramètres du modèle LFMM par maximisation de la vraisemblance consiste à minimiser l'opposé du logarithme de la vraisemblance des paramètres du modèle LFMM, que l'on peut écrire de la manière suivante :

$$L(U,V,B,\sigma^2) = \frac{1}{2\sigma^2}||G - UV^T - XB^T||_F^2 + \frac{nL}{2}\ln(2\pi\sigma^2) \tag{6.1}$$

6.2.1.2 Une approche par maximisation de la vraisemblance régularisée

Une approche régularisée d'estimation des paramètres du modèle LFMM par maximisation de la vraisemblance régularisée consiste à minimiser l'opposé du logarithme de la vraisemblance des paramètres du modèle LFMM auquel on ajoute une pénalité. Ici, nous avons choisi une pénalité de type "Ridge". On parlera d'approche fréquentiste régularisée de type "Ridge". Dans un cadre bayésien, cette fonction correspond à l'opposé du logarithme de la loi a posteriori de B où les vecteurs B_ℓ ont, pour loi a priori, une loi normale $N(0, D_B)$. La matrice D_B de taille $d \times d$ est diagonale. Cette matrice D_B est un paramètre que l'on ne cherchera pas à estimer par cette approche. La fonction à minimiser s'écrit

$$L_r(U,V,B,\sigma^2) = \frac{1}{2\sigma^2}||G-UV^T-XB^T||_F^2+\frac{1}{2}tr(B^TD_B^{-1}B)+\frac{nL}{2}\ln(2\pi\sigma^2)+\frac{L}{2}\ln(2\pi|D_B|) \tag{6.2}$$

6.2.2 Des approches probabilistes

Dans le cadre d'approches probabilistes, on suppose que les coefficients de régression du modèle LFMM, B_ℓ, suivent a priori une loi normale $N(0, \sigma^2 D_B)$ pour le locus ℓ et que les éléments de la matrice V suivent, a priori, une loi normale $N(0, 1)$. Dans cette approche, les variables U, σ^2 et D_B sont des paramètres inconnus et B et V des variables aléatoires.

On peut, alors, moyenner la vraisemblance selon V afin d'obtenir la loi conditionnelle de B en fonction des paramètres U, σ^2 et D_B. Cette loi s'écrit

$$B_\ell | G, U, D_B, \sigma^2 \sim N(m_{B\ell}, \Sigma_B) \quad \ell = 1 \dots L \tag{6.3}$$

où

$$\Sigma_B = (\frac{D_B^{-1}}{\sigma^2} + X^T (UU^T + \sigma^2 I_n)^{-1} X)^{-1}$$

$$m_{B\ell} = \Sigma_B X^T (UU^T + \sigma^2 I_n)^{-1} G_\ell$$

Nous proposons deux approches, similaires aux approches proposées par Tipping and Bishop (1999), afin d'obtenir les estimateurs de maximum de vraisemblance de U, σ^2 et D_B : un algorithme EM et une maximisation de la vraisemblance moyennée selon V et B. Une fois que nous avons obtenu les estimateurs de maximum de vraisemblance pour U, σ^2 et D_B, nous pouvons utiliser la formule (6.3), afin de calculer la loi conditionnelle de B sachant les estimateurs EM de U, σ^2 et D_B.

Pour rappel, le modèle appelé "Probabilistic PCA" proposé par Tipping and Bishop (1999) s'écrit

$$G_{i\ell} = U_i V_\ell^T + \epsilon_{i\ell} \quad i = 1 \dots n, \quad \ell = 1 \dots L,$$

où la variable aléatoire V_ℓ a pour loi a priori une loi normale $N(0, I_K)$ pour chaque locus ℓ, $\epsilon_{i\ell}$ est un résidu de loi $N(0, \sigma^2)$ et la matrice U des vecteurs $(U_i)_i$ est un paramètre inconnu.

6.2.2.1 Un algorithme EM

Nous proposons un algorithme EM pour l'estimation des paramètres U, σ^2 et D_B. Pour cela nous appliquons un algorithme EM à la log-vraisemblance complète, moyennée selon

B, qui s'écrit de la manière suivante :

$$L_c(V, U, \sigma^2, D_B) = \sum_{\ell} \ln(p(G_\ell, V_\ell | U, D_B, \sigma^2))$$

où

$$p(V_\ell, G_\ell | U, D_B, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2} |\Sigma_X|^{1/2}} \exp(-\frac{1}{2\sigma^2}(G_\ell - UV_\ell^T)^T \Sigma_X^{-1}(G_\ell - UV_\ell^T))$$

$$\times \frac{1}{(2\pi)^{K/2}} \exp(-\frac{||V_\ell||^2}{2})$$

et $\Sigma_X = XD_BX^T + I_n$.

6.2.2.2 Une approche par maximisation de vraisemblance marginale

Une autre approche consiste à moyenner la vraisemblance selon les variables aléatoires V et B afin d'obtenir une vraisemblance marginale. On cherche ainsi les valeurs de U, σ^2 et D_B maximisant la log-vraisemblance marginale, qui s'écrit de manière suivante :

$$L_m(U, D_B, \sigma^2) = -\frac{L}{2}(n \ln(2\pi) + \ln|C| + tr(C^{-1}S))$$

où $C = UU^T + \sigma^2(I_n + XD_BX^T)$ et $S = \frac{GG^T}{L}$.

6.2.3 Des approches bayésiennes

Dans le cadre d'approches bayésiennes, nous considèrons que les coefficients de régression du modèle LFMM, B_ℓ, suivent a priori une loi normale $N(0, D_B)$ pour le locus ℓ, que les éléments de la matrice V suivent une loi normale $N(0, 1)$ et que les éléments de la matrice U suivent une loi normale $N(0, \sigma_U^2)$. Les paramètres sont D_B, σ^2 et σ_U^2. Les variables aléatoires sont les matrices B, V et U, et nous cherchons à obtenir les lois a posteriori des ces variables, ou des estimations ponctuelles de ces variables.

6.2.3.1 Une approche par maximisation de la loi a posteriori

On peut chercher les valeurs de U, V et B minimisant l'opposé du logarithme de la loi a posteriori L_{MAP} qui s'écrit de la manière suivante :

$$L_{MAP}(U, V, B, \sigma^2, D_B, \sigma_U^2) = \frac{1}{2\sigma^2}||G - UV^T - XB^T||_F^2 + \frac{1}{2\sigma_U^2}||U||_F^2 + \frac{1}{2}||V||_F^2 + \frac{1}{2}tr(BD_B^{-1}B^T)$$

$$+\frac{nL}{2}\ln(2\pi\sigma^2)+\frac{nK}{2}\ln(2\pi\sigma_U^2)+\frac{dL}{2}\ln(2\pi)+\frac{L}{2}\ln(|D_B|).$$

6.2.3.2 Une approche par inférence variationnelle bayésienne

L'approche par inférence variationnelle consiste a approcher les loi a posteriori par des lois connues. Une description de l'algorithme Variational Bayes est similaire à celle proposée par Nakajima et al. (2011).

La loi a posteriori du modèle de LFMM s'écrit

$$p(U,V,B|G,\sigma^2,D_B,\sigma_U^2)=\frac{p(G|U,V,B,\sigma^2)p(U|\sigma_U^2)p(V)p(B|D_B)}{Z(G)}$$

où $Z(G)=E(p(G|U,V,B,\sigma^2))_{p(U)p(V)p(B)}$ est la vraisemblance marginale, et où on note $E(.)_p$ l'espérance sachant la loi p. Comme la distribution de la loi a posteriori n'est pas calculable, on peut proposer une approche variationnelle. On pose $r(U,V,B)$, aussi notée r, une loi cible. On cherche à approcher la loi a posteriori par la loi cible. Pour cela, on définit la fonction suivante, comme étant l'énergie libre,

$$F(r|G)=E(\ln\frac{r(U,V,B)}{p(G|U,V,B,\sigma^2)p(U|\sigma_U^2)p(V)p(B|D_B)})_{r(U,V,B)}$$

$$=E(\ln\frac{r(U,V,B)}{p(U,V,B|G,\sigma^2,D_B,\sigma_U^2)})_{r(U,V,B)}-\ln Z(G)$$

Dans cette équation, le premier terme est la divergence de Kullback-Leibler entre la loi cible et la loi a posteriori et le second terme est une constante. Par conséquent, pour minimiser l'énergie libre, il suffit de déterminer la loi la plus proche de la loi a posteriori au sens de la divergence de Kullback-Leibler.

6.2.3.3 Une approche par échantillonnage de Gibbs

L'échantillonnage de Gibbs consiste à obtenir une estimation des lois a posteriori par échantillonnage selon les lois conditionnelles. L'échantillonnage de Gibbs est l'algorithme utilisé actuellement pour l'estimation des paramètres du modèle de LFMM. Pour cet algorithme, on suppose que σ^2, σ_U^2 et D_B sont des variables aléatoires qui suivent, a priori, des lois inverse-Gamma $\Gamma^{-1}(\eta,\eta)$ où nous avons choisi $\eta=10^2-10^3$.

6.3 Approches fréquentistes

Nous proposons deux approches fréquentistes pour estimer les paramètres du modèle LFMM. La première approche est une approche par maximisation de la vraisemblance. Nous verrons que l'estimateur de la matrice des coefficients de régression n'est pas unique. Nous proposons donc une approche par maximisation de la vraisemblance régularisée. Dans ce cas, l'estimateur des coefficients de régression du modèle LFMM est bien défini.

6.3.1 Estimateurs de maximum de vraisemblance

L'opposé de la log-vraisemblance des paramètres du modèle de LFMM s'écrit

$$L(U, V, B, \sigma^2) = \frac{1}{2\sigma^2}||G - UV^T - XB^T||_F^2 + \frac{nL}{2}\ln(2\pi\sigma^2). \tag{6.4}$$

Résultat : Les valeurs B_v, de la matrice des coefficients de régression, maximisant la vraisemblance du modèle de LFMM sont de la forme

$$B_v^T = (X^T X)^{-1} X^T G - C_v V_v^T \tag{6.5}$$

où la matrice C_v est de taille $D \times K$ et quelconque et la matrice V_v est la matrice unitaire des poids de la décomposition en valeurs singulières de rang K, $U_{svd} D_{svd} {V_v}^T = svd_K(P_{X^\perp} G)$. Il n'y a donc pas unicité de la solution pour B_v.

L'estimateur σ_v^2 de la variance résiduelle est l'estimateur empirique de la variance du résidu. Comme les estimateurs B_v et V_v ne dépendent pas de σ_v^2, nous ne nous intéressons pas à cet estimateur.

Idées de la démonstration :

On peut écrire le système aux dérivées partielles associé à la fonction L de la manière suivante :

$$\begin{cases} \dfrac{\partial L}{\partial U} = \dfrac{1}{\sigma^2}(UV^T + XB^T - G)V \\ \dfrac{\partial L}{\partial V} = \dfrac{1}{\sigma^2}U^T(UV^T + XB^T - G) \\ \dfrac{\partial L}{\partial B} = \dfrac{1}{\sigma^2}X^T(UV^T + XB^T - G) \\ \dfrac{\partial L}{\partial \sigma^2} = -\dfrac{1}{2(\sigma^2)^2}||G - UV^T - XB^T||_F^2 + \dfrac{nL}{2\sigma^2} \end{cases} \tag{6.6}$$

Ce système d'équations s'annule pour B_v, U_v et V_v lorsque

$$\begin{cases} {B_v}^T = (X^TX)^{-1}X^T(G - U_v{V_v}^T) \\ 0 = {U_v}^T P_{X^\perp}(G - U_v{V_v}^T) \\ 0 = P_{X^\perp}(G - U_v{V_v}^T)V_v \end{cases} \tag{6.7}$$

On décompose U_v et G selon les sous-espaces supplémentaires $vect(X)$ et $vect(X^\perp)$

$$U_v = U_X^v + U_{X^\perp}^v$$

$$G = G_X + G_{X^\perp}$$

et on reprend le système d'équations ci-dessus

$$\begin{cases} {B_v}^T = (X^TX)^{-1}X^TG - (X^TX)^{-1}X^T(U_X^v{V_v}^T) \\ 0 = {U_{X^\perp}^v}^T(G_{X^\perp} - U_{X^\perp}^v{V_v}^T) \\ 0 = (G_{X^\perp} - U_{X^\perp}^v{V_v}^T)V_v \end{cases} \tag{6.8}$$

On remarque que la seconde et la troisième équations forment un sous-système pour lequel

$$U_{X^\perp}^v{V_v}^T = svd_K(P_{X^\perp}G).$$

Sans restriction, on pourra supposer que V_v est une matrice unitaire. Dans ce cas, V_v est donnée par la solution de la décomposition en valeurs singulières, $U_{svd}D_{svd}{V_v}^T = svd_K(P_{X^\perp}G)$.

De plus, le système n'impose pas de contrainte sur la matrice $(X^TX)^{-1}X^TU_X^v$ que l'on notera C_v. On a donc pour tout C_v de taille $d \times K$,

$${B_v}^T = (X^TX)^{-1}X^TG - C_v{V_v}^T$$

où V_v a une unique solution, donnée par la solution de la svd $U_{svd}D_{svd}{V_v}^T = svd_K(P_{X^\perp}G)$.

Comme on a pu le remarquer précédemment, la solution B^v n'est pas unique car $P_{X^\perp}$ n'est pas inversible. Parmi ces estimateurs, on retrouve l'estimateur de la régression linéaire dans le cas où $C_v = 0$, c'est-à-dire dans le cas où U_v et X sont orthogonaux. Cela correspond au cas où il n'y a pas de variable non observée corrélée à X. Dans le cas $C_v \neq 0$, on remarque, intuitivement, que le but du modèle LFMM est de corriger l'estimateur de la régression linéaire $(X^TX)^{-1}X^TG$ par la projection, dans l'espace engendré par X, des facteurs de confusion, $C_v{V_v}^T$.

6.3.2 Estimateur régularisé des coefficients de régression

Comme on a pu le voir, il existe plusieurs estimateurs des coefficients de régression maximisant la vraisemblance. Ceci est dû au fait que $P_{X^\perp}$ n'est pas inversible. Une façon de contourner cette contrainte est d'utiliser une matrice projection oblique $P^D_{X^\perp}$. Ceci revient à pénaliser la vraisemblance par une pénalité de type "Ridge".

La vraisemblance pénalisée s'écrit

$$L_r(U,V,B,\sigma^2) = \frac{1}{2\sigma^2}||G-UV^T-XB^T||^2_F+\frac{1}{2}tr(B^TD_B^{-1}B)+\frac{nL}{2}\ln(2\pi\sigma^2)+\frac{L}{2}\ln(2\pi|D_B|) \tag{6.9}$$

Nous considérons que σ^2 et D_B sont des paramètres fixes que l'on ne cherchera pas à estimer par maximum de vraisemblance pénalisée. On remarque qu'en fait, seul le rapport $\frac{D_B}{\sigma^2}$, intervient dans l'estimation de B, V et U. On notera ce ratio $\tilde{D} = \frac{D_B}{\sigma^2}$. On considère que le choix de ce rapport fait partie du choix de modèle. On peut choisir la valeur de ce rapport par validation croisée.

Résultat : Les valeurs U_r, V_r et B_r de U, V et B minimisant L_r sont de la forme

$$B_r^T = (X^TX)^{-1}X^T(G-U_rV_r^T) \tag{6.10}$$

avec

$$U_rV_r^T = C_h^{-1}svd_K(C_hG) \tag{6.11}$$

où C_h est la matrice de décomposition de Cholesky de $P^{\tilde{D}}_{X^\perp}$, $P^{\tilde{D}}_{X^\perp} = C_h^TC_h$.

Idées de démonstration : Le système aux dérivées partielles associé à L_r s'annule pour

$$\begin{cases} B_r^T = (X^TX+\tilde{D}^{-1})^{-1}X^T(G-U_rV_r^T) \\ 0 = U_r^TP^{\tilde{D}}_{x^\perp}(G-U_rV_r^T) \\ 0 = P^{\tilde{D}}_{x^\perp}(G-U_rV_r^T)V_r \end{cases} \tag{6.12}$$

La seconde et la troisième équation du système 6.12 forment un sous système pour lequel

$$U_rV_r^T = C_h^{-1}svd_K(C_hG).$$

On remarque que les solutions de U et V ne sont pas uniques du fait de la non unicité de la décomposition en facteurs. Toutefois, la solution de B est unique puisqu'elle ne dépend que de la solution UV^T qui est unique. L'estimateur Ridge des coefficients de régression du modèle LFMM est donc bien défini. Cela provient du fait que l'on peut inverser une matrice de projection oblique.

6.4 Approche probabiliste

Dans la section précédente, nous avons proposé des estimateurs du coefficient de régression B fondés sur la maximisation de la vraisemblance et de la vraisemblance pénalisée. Ces approches ne considèrent pas de régularisation sur l'estimation des facteurs. Nous proposons donc un modèle probabiliste, extension du modèle "probabilistic PCA" proposé par Tipping and Bishop (1999), prenant en compte une régularisation dans l'estimation des facteurs.

Ce modèle probabiliste peut s'écrire de la manière suivante

$$G_{i\ell} = U_i V_\ell^T + X_i B_\ell^T + \epsilon_{i\ell}$$

avec $\epsilon_{i\ell}$ un résidu de loi $N(0, \sigma^2)$, V de loi a priori $N(0, I_K)$ et B de loi a priori $N(0, \sigma^2 D_B)$ où D_B est une matrice diagonale de taille $d \times d$. On remarque que que la matrice de covariance a priori de B est $\sigma^2 D_B$ (plutôt que D_B).

On peut réécrire ce modèle, en intégrant la vraisemblance selon B, de la manière suivante

$$G_{i\ell} = U_i V_\ell^T + \epsilon_{i\ell} \tag{6.13}$$

où ϵ_ℓ est un vecteur $(\epsilon_{i\ell})_i$ de loi $N(0, \sigma^2(I_n + X D_B X^T))$ et V_ℓ suit la loi $N(0, I_K)$.

La loi conditionnelle de B sachant U, σ^2 et D_B (et intégrée selon V) s'écrit

$$B_\ell | G, U, D_B, \sigma^2 \sim N(m_{B\ell}, \Sigma_B) \quad \ell = 1 \dots L \tag{6.14}$$

où

$$\Sigma_B = (\frac{D_B^{-1}}{\sigma^2} + X^T (UU^T + \sigma^2 I_n)^{-1} X)^{-1}$$

$$m_{B\ell} = \Sigma_B X^T (UU^T + \sigma^2 I_n)^{-1} G_\ell.$$

L'objectif est donc de calculer des estimateurs EM de U, σ^2 et D_B, afin d'obtenir la loi conditionnelle de B sachant les estimateurs EM de U, σ^2 et D_B.

6.4.1 Algorithme EM pour LFMM

On peut écrire l'algorithme EM correspondant à l'équation (6.13), Pour cela, on écrit l'algorithme EM du modèle probabilistic PCA (Tipping and Bishop, 1999) pour lequel on considère une matrice de covariance $\Sigma = \sigma^2(X D_B X^T + I_n)$ pour le bruit. On note

$\Sigma_X = XD_BX^T + I_n$. On écrit la log-vraisemblance complète du modèle

$$L_C = \sum_\ell \ln(p(G_\ell, V_\ell|U, D_B, \sigma^2)),$$

où

$$p(V_\ell, G_\ell|U, D_B, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}|\Sigma_X|^{1/2}} \exp(-\frac{1}{2\sigma^2}(G_\ell - UV_\ell^T)^T\Sigma_X^{-1}(G_\ell - UV_\ell^T))$$

$$\times \frac{1}{(2\pi)^{K/2}} \exp(-\frac{||V_\ell||^2}{2})$$

Pour l'étape E, on considère l'espérance de L_C suivant la loi $p(V_\ell|G_\ell, U, \sigma^2, D_B)$:

$$E(L_C) = -\frac{1}{2}\sum_\ell [\ln|\Sigma_X| + n\ln\sigma^2 + tr(E(V_\ell^T V_\ell)) + \frac{1}{\sigma^2}G_\ell^T\Sigma_X^{-1}G_\ell - \frac{2}{\sigma^2}E(V_\ell)U^T\Sigma_X^{-1}G_\ell$$

$$+\frac{1}{\sigma^2}tr(U^T\Sigma_X^{-1}UE(V_\ell^T V_\ell))]$$

où l'on a omit les termes indépendants des paramètres et où

$$E(V_\ell^T) = (U^T\Sigma_X^{-1}U + I_K)^{-1}U^T\Sigma_X^{-1}G_\ell$$

et

$$E(V_\ell^T V_\ell) = \sigma^2(U^T\Sigma_X^{-1}U + I_K)^{-1} + E(V_\ell)E(V_\ell)^T.$$

Pour l'étape M de maximisation, on maximise l'espérance $E(L_C)$ par rapport à U et σ^2. On obtient comme maximum pour U,

$$U_{em} = [\sum_\ell G_\ell E(V_\ell)][\sum_\ell E(V_\ell^T V_\ell)]^{-1}$$

et pour σ^2,

$$\sigma^2_{em} = \frac{1}{nL}\sum_\ell [G_\ell^T\Sigma_X^{-1}G_\ell - 2E(V_\ell)U^T\Sigma_X^{-1}G_\ell + tr(U^T\Sigma_X^{-1}UE(V_\ell^T V_\ell))].$$

On remarque qu'il est plus difficile d'obtenir une formule analytique de l'estimateur EM de D_B. Une possibilité serait de calculer la dérivée partielle de $E(L_C)$ par rapport à D_B puis de déterminer une valeur de D_B qui annule cette dérivée par un algorithme de descente de gradient par exemple.

Algorithme : L'algorithme (EM) de ce modèle pour une valeur fixée de D_B consiste à mettre à jour alternativement les paramètres du modèle par les règles suivantes

$$\begin{cases} E(V_\ell^T) & \leftarrow & (U_{em}^T \Sigma_X^{-1} U_{em} + I_K)^{-1} U_{em}^T \Sigma_X^{-1} G_\ell \\ E(V_\ell^T V_\ell) & \leftarrow & \sigma_{em}^2 (U_{em}^T \Sigma_X^{-1} U_{em} + I_K)^{-1} + E(V_\ell) E(V_\ell)^T \\ U_{em} & \leftarrow & [\sum_\ell G_\ell E(V_\ell)][\sum_\ell E(V_\ell^T V_\ell)]^{-1} \\ \sigma_{em}^2 & \leftarrow & \frac{1}{nL} \sum_\ell [G_\ell^T \Sigma_X^{-1} G_\ell - 2E(V_\ell) U_{em}^T \Sigma_X^{-1} G_\ell + tr(U_{em}^T \Sigma_X^{-1} U_{em} E(V_\ell^T V_\ell))] \\ \text{mise à jour de } D_B. \end{cases} \tag{6.15}$$

6.4.2 Estimateur de la vraisemblance marginale

De manière similaire au raisonnement de Tipping and Bishop (1999), on peut écrire la vraisemblance marginale en intégrant la vraisemblance selon les variables V et B. La log-vraisemblance marginale s'écrit de la manière suivante :

$$L_m(U, D_B, \sigma^2) = -\frac{L}{2}(n \ln(2\pi) + \ln|C| + tr(C^{-1}S)) \tag{6.16}$$

où $S = \frac{GG^T}{L}$ et $C = UU^T + \sigma^2(I_n + X D_B X^T)$.

On pose alors $I_n + X D_B X^T = P_X D_X^2 P_X^T$ où $P_X D_X^2 P_X^T$ est la décomposition en valeurs singulières de $I_n + X D_B X^T$. Pour remarque, D_X est une matrice diagonale, de diagonale non nulle. On pose

$$\tilde{U} = D_X^{-1} P_X^T U$$

$$\tilde{S} = D_X^{-1} P_X^T S P_X D_X^{-1}$$

$$\tilde{C} = \tilde{U}\tilde{U}^T + \sigma^2 I_n.$$

On peut alors écrire L_m comme

$$L_m(U, D_B, \sigma^2) = -\frac{L}{2}(n \ln(2\pi) + \ln|\tilde{C}| + \ln D_X^2 + tr(\tilde{C}^{-1}\tilde{S})) \tag{6.17}$$

Tipping and Bishop (1999) ont montré que les valeurs de $\tilde{U}$ et σ^2 maximisant L_m sont

$$\tilde{U}_m = U_{\tilde{S}_K} (D_{\tilde{S}_K} - \sigma_m^2 I_n)^{1/2} R$$

où $U_{\tilde{S}_K}$ est la matrice des K premiers vecteurs propres de $\tilde{S}$, R est une matrice orthogonale de taille $K \times K$ quelconque, et

$$D_{\tilde{S}_K}(j) = \begin{cases} \lambda_j \text{ si } j \leq K \\ \sigma_m^2 \text{ sinon,} \end{cases} \tag{6.18}$$

où λ_j est la $j$ème valeur propre de $\tilde{S}$ et

$$\sigma_m^2 = \frac{1}{n-K} \sum_{j=K+1}^{n} \lambda_j.$$

Résultat : Les valeurs de U et σ^2 maximisant L_m sont

$$\sigma_m^2 = \frac{1}{n-K} \sum_{j=K+1}^{n} \lambda_j$$

où λ_j est la $j$ème valeur propre de $\tilde{S}$ et

$$U_m = P_X D_X U_{\tilde{S}_K} (D_{\tilde{S}_K} - \sigma_m^2 I_n)^{1/2} R.$$

On peut alors écrire à la fonction L_m avec les valeurs maximales U_m et σ_m^2 sous la forme

$$L_m(U_m, D_B, \sigma_m^2) = -\frac{L}{2}\{\sum_{j=1}^{K} \ln(\lambda_j) - (n-K)\ln(\frac{1}{n-K}\sum_{j=K+1}^{n} \lambda_j) - n\ln(2\pi) + K + \ln D_X^2\}$$

L'écriture de la fonction L_m dépend de D_B à travers D_X et $(\lambda_j)_j$ les valeurs propres de $\tilde{S}$. On peut donc imaginer déterminer une valeur de D_B maximisant la fonction L_m conditionnellement à U_m et σ_m^2 par une recherche exhaustive.

6.5 Approche bayésienne

L'approche bayésienne, pour LFMM, est fondée sur une correction des facteurs de confusion à partir de la factorisation de matrice bayésienne (Frichot et al., 2013). Le modèle est décrit ci-dessous.

$$G_{i\ell} = U_i V_\ell^T + X_i B_\ell^T + \epsilon_{i\ell},$$

où B_ℓ est le vecteur des coefficients de regression de dimension d, X_i est un vecteur d'indicateurs environnementaux de dimension d. Nous supposons que B_ℓ est de loi a priori $N(0, D_B)$ avec D_B une matrice de covariance diagonale. Les termes U_i et V_ℓ sont

des vecteurs de dimension K modélisant les facteurs de confusion. Le terme U_i est de loi a priori $N(0, \sigma_U^2 I_K)$ et le terme V_ℓ est de loi a priori $N(0, I_K)$. Le terme $\epsilon_{i\ell}$ est un terme de résidu de loi a priori $N(0, \sigma^2)$. Contrairement à l'approche probabiliste, dans l'approche bayésienne, U est une variable aléatoire et non un paramètre.

6.5.1 Maximum A Posteriori (MAP)

L'opposé du logarithme de la loi a posteriori du modèle de LFMM s'écrit de la manière suivante :

$$L_{MAP}(U, V, B, \sigma^2, D_B, \sigma_U^2) = \frac{1}{2\sigma^2}||G - UV^T - XB^T||_F^2 + \frac{1}{2\sigma_U^2}||U||_F^2 + \frac{1}{2}||V||_F^2 + \frac{1}{2}tr(BD_B^{-1}B^T)$$

$$+ \frac{nL}{2}\ln(2\pi\sigma^2) + \frac{nK}{2}\ln(2\pi\sigma_U^2) + \frac{dL}{2}\ln(2\pi) + \frac{L}{2}\ln(|D_B|).$$

On peut écrire le système aux dérivées partielles associé à L_{MAP} de la manière suivante :

$$\begin{cases} \frac{\partial L_{MAP}}{\partial B^T} = \frac{1}{\sigma^2}X^T(UV^T + XB^T - G) + D_B^{-1}B^T \\ \frac{\partial L_{MAP}}{\partial V^T} = \frac{1}{\sigma^2}U^T(UV^T + XB^T - G) + V^T \\ \frac{\partial L_{MAP}}{\partial U} = \frac{1}{\sigma^2}(UV^T + XB^T - G)V + \frac{1}{\sigma_U^2}U \\ \frac{\partial L_{MAP}}{\partial \sigma^2} = -\frac{1}{2(\sigma^2)^2}||UV^T + XB^T - G||_F^2 + \frac{nL}{2\sigma^2} \\ \frac{\partial L_{MAP}}{\partial \sigma_U^2} = -\frac{1}{2(\sigma_U^2)^2}||U||_F^2 + \frac{nK}{2\sigma_U^2} \\ \frac{\partial L_{MAP}}{\partial D_B(j)} = -\frac{1}{2D_B(j)^2}||B_j||_F^2 + \frac{L}{2D_B(j)} \end{cases} \tag{6.19}$$

Algorithme : On peut alors en déduire un algorithme itératif pour minimiser localement L_{MAP} en itérant les six règles suivantes :

$$\begin{cases} U^T \leftarrow (V^T V + \dfrac{\sigma^2}{\sigma_U^2} I_K)^{-1} V^T (G - XB^T)^T \\ V^T \leftarrow (U^T U + \sigma^2 I_K)^{-1} U^T (G - XB^T) \\ B^T \leftarrow (X^T X + \sigma^2 D_B^{-1})^{-1} X^T (G - UV^T) \\ \sigma_U^2 \leftarrow \dfrac{||U||_F^2}{nK} \\ \sigma^2 \leftarrow \dfrac{||G - UV^T - XB^T||_F^2}{nL} \\ D_B(j) \leftarrow \dfrac{||B_j||_F^2}{L}, \quad j = 1 \dots d. \end{cases} \tag{6.20}$$

On peut voir cet algorithme comme un algorithme de minimisation des moindres carrées alternées selon B, V et U.

Cependant, comme pointé par Raiko et al. (2007) et Nakajima et al. (2011), les estimateurs de σ_U^2, σ^2 et de D_B ne fonctionnent pas en pratique car la fonction a minimiser n'est pas bornée lorsque σ_U^2, σ^2 et D_B tendent vers 0. Cela correspond à une solution triviale à moins que l'algorithme itératif s'arrête lorsqu'il a atteint un minimum local quelconque.

6.5.2 Algorithme Variational Bayes (VB)

Cette description de l'algorithme Variational Bayes est similaire à celle proposée par Nakajima et al. (2011).

Comme expliqué précedemment, l'objectif de l'approche bayésienne variationnelle est d'approcher la loi a posteriori $p(U, V, B|G, \sigma^2, D_B, \sigma_U^2)$ et une loi cible $r(U, V, B)$. Pour cela, on cherche à minimiser la distance entre la loi a posteriori et la loi cible à l'aide de la divergence de Kullback-Leibler. La fonction à minimiser s'écrit de la manière suivante :

$$KL(r|G) = E(\ln \frac{r(U, V, B)}{p(U, V, B|G, \sigma^2, D_B, \sigma_U^2)})_{r(U,V,B)}$$

Dans cette équation, le premier terme est la divergence de Kullback-Leibler entre la loi cible et la loi a posteriori et le second terme est une constante. Par conséquent, pour minimiser l'énergie libre, il suffit de déterminer la loi la plus proche de la loi a posteriori au sens de la divergence de Kullback-Leibler.

On suppose indépendance entre les matrices U,V et B. Cela se traduit par

$$r^{VB}(U,V,B) = r_U^{VB}(U) r_V^{VB}(V) r_B^{VB}(B).$$

Cette hypothèse permet d'écrire un algorithme itératif calculable. De plus, sous cette hypothèse, la loi cible minimisant l'énergie libre s'écrit

$$r^{VB}(U,V,B) = \prod_{i=1}^{n} N(U_i; \hat{U}_i, \Sigma_U) \prod_{l=1}^{L} N(V_l; \hat{V}_l, \Sigma_V) N(B_l; \hat{B}_l, \Sigma_B)$$

où $N(x; \hat{x}, \Sigma_x)$ est la densité de loi normale multivariée en x de moyenne $\hat{x}$ et de matrice de covariance Σ_x.

Les paramètres de la loi cible sont

$$\hat{U} = (G - XB^T)\hat{V}\frac{\Sigma_U}{\sigma^2}$$

$$\Sigma_U = \sigma^2(\hat{V}^T\hat{V} + L\Sigma_V + \frac{\sigma^2}{\sigma_U^2} I_K)^{-1}$$

$$\hat{V} = (G - XB^T)^T\hat{U}\frac{\Sigma_V}{\sigma^2}$$

$$\Sigma_V = \sigma^2(\hat{U}^T\hat{U} + n\Sigma_U + \sigma^2 I_K)^{-1}$$

$$\hat{B} = (G - UV^T)^T X\frac{\Sigma_B}{\sigma^2}$$

$$\Sigma_B = \sigma^2(X^T X + \sigma^2 D_B^{-1})^{-1}$$

Algorithme : On peut dont mettre à jour itérativement $\hat{U}, \Sigma_U, \hat{V}, \Sigma_V, \hat{B}$ et Σ_B jusqu'à converger vers une minimum local de l'énergie libre.

6.5.3 Approximation Variational Bayes simple

Une version simplifié de l'approximation bayésienne variationnelle suppose l'indépendance entre les colonnes de chaque matrice $(U_k)_{k=1..K}$, $(V_k)_{k=1..K}$ et $(B_j)_{j=1..d}$

$$r^{simpleVB}(U,V,B) = \prod_{k=1}^{K} r_{U_k}^{simpleVB}(U_k) r_{V_k}^{simpleVB}(V_k) \prod_{j=1}^{d} r_{B_j}^{simpleVB}(B_j)$$

Cette contrainte restreint les covariance Σ_U, Σ_V, et Σ_B à être diagonale.

La loi cible s'écrit alors

$$r^{simpleVB}(U,V,B) = \prod_{k=1}^{K} N(U_k; \hat{U}_k, \sigma^2_{U_k} I_n) N(V_k; \hat{V}_k, \sigma^2_{V_k} I_L) \prod_{j=1}^{d} N(B_j; \hat{B}_j, \sigma^2_{B_j} I_L)$$

où les paramètres vérifient pour, $k = 1 \dots K$ et $j = 1 \dots d$,

$$\hat{U}_k = \frac{\sigma^2_{U_k}}{\sigma^2}(G - XB^T - \sum_{k' \neq k} \hat{U}_{k'} \hat{V}_{k'}^T)\hat{V}_k$$

$$\sigma^2_{U_k} = \sigma^2(||\hat{V}_k||_F^2 + L\sigma^2_{V_k} + \frac{\sigma^2}{\sigma^2_U})^{-1}$$

$$\hat{V}_k = \frac{\sigma^2_{V_k}}{\sigma^2}(G - XB^T - \sum_{k' \neq k} \hat{U}_{k'} \hat{V}_{k'}^T)^T \hat{U}_k$$

$$\sigma^2_{V_k} = \sigma^2(||\hat{U}_k||_F^2 + n\sigma^2_{U_k} + \sigma^2)^{-1}$$

$$\hat{B}_j = \frac{\sigma^2_{B_j}}{\sigma^2}(G - UV^T - \sum_{j' \neq j} \hat{X}_{j'} \hat{B}_{j'}^T)^T X_j$$

$$\sigma^2_{B_j} = \sigma^2(||X_j||_F^2 + \frac{\sigma^2}{\sigma^2_{B_j}})^{-1}$$

Algorithme : On peut donc itérer les égalités ci-dessus pour obtenir un minimum local de l'énergie libre.

On peut montrer par un raisonnement similaire à Nakajima et al. (2011) que des matrices diagonales pour Σ_U et Σ_V minimisent l'énergie et qu'une matrice diagonale pour Σ_B minimise l'énergie si et seulement si $X^T X$ est une matrice diagonale. On pourra donc considérer un algorithme semi-simple de Variational Bayes où l'on considérera seulement l'indépendance entre colonnes des matrices U et V.

6.6 Discussion sur les estimateurs du modèle LFMM

La méthode actuellement utilisée dans le logiciel `LFMM` pour l'estimation des paramètres du modèle LFMM est l'échantillonnage de Gibbs. Cet algorithme permet d'obtenir une estimation de la loi a posteriori des coefficients de régression et de pouvoir effectuer un test. Comme expliqué précédemment, il présente aussi certaines limites. Tout d'abord, l'algorithme peut être coûteux en calcul. De plus, il est nécessaire de faire un certain

nombre de choix de modélisation, comme celui des lois a priori (Zhou et al., 2013). Actuellement, les lois a priori sur U, V et B sont des lois normales et les loi a priori sur les variances sont des lois inverse-gamma. On pourrait utiliser, par un exemple, un mélange de lois gaussiennes comme loi a priori sur les coefficients de régression afin de modéliser les effets polygéniques (Zhou et al., 2013). Enfin, en grande dimension, il se peut que l'algorithme de Gibbs parcourt seulement un sous ensemble de l'espace d'état dépendant de l'initialisation (West, 2003). Il serait donc intéressant de considérer une initialisation adéquate de l'échantillonneur de Gibbs.

Nous avons présenté une formule explicite pour les estimateurs par maximum de vraisemblance des paramètres du modèle de LFMM. Intuitivement, cette solution nous permet de voir qu'il existe deux manières équivalentes de corriger la régression linéaire en utilisant un modèle à facteurs. On peut, soit corriger l'estimation du coefficient de régression en estimant le coefficient de régression du modèle de LFMM et en testant l'hypothèse "$B_{LFMM} = 0$", soit modifier le test de la régression linéaire en testant "$B_{LM} = VC^T$". Toutefois ce test est difficile à réaliser car les matrices V et C sont inconnues.

L'estimateur de type Ridge proposé pour les coefficients de régression de LFMM permet d'obtenir un estimateur de maximum de vraisemblance régularisé global et unique des coefficients de régression. On souhaite, ensuite, effectuer un test associé à chaque coefficient de régression. Une idée serait d'obtenir une estimation de la loi de B_r sous l'hypothèse nulle en effectuant des permutations aléatoires ou un algorithme de bootstrap (Chung and Storey, 2013), afin de comparer la valeur obtenue de B_r à la loi nulle sous H_0. Toutefois, on réalise qu'il faudrait effectuer des millions de permutations pour obtenir une loi de test pour B_r, suffisament précise pour prendre en compte la correction pour le test multiple. Alternativement, on peut chercher à obtenir un estimateur de la variance de B_r et faire une hypothèse sur la loi associée au test. De plus, l'estimation ponctuelle obtenue par cet algorithme peut être un point d'initialisation cohérent de l'échantillonneur de Gibbs. Une limite de l'estimateur Ridge est qu'il dépend de la valeur du paramètre de régularisation utilisée. Toutefois, on peut s'attendre à ce que ce paramètre de régularisation ait des interprétations similaires au paramètre de régularisation du problème de type Ridge de la régression linéaire (Tikhonov, 1943; Hoerl and Kennard, 1970). On peut le choisir a posteriori grâce à un critère prédictif tel que la validation croisée ou l'entropie croisée. Nous avons choisi une régularisation de type Ridge car cela nous permet d'obtenir le maximum global de la vraisemblance pénalisée. Toutefois, il existe de nombreux autres possibilités de régularisation comme la régularisation Lasso (Tibshirani, 1996). Une approche de type "Lasso" pourrait apporter de la parcimonie dans l'estimation des coefficients de régression et ainsi aider à résoudre le fléau de la grande dimension.

L'algorithme EM proposé apporte un algorithme itératif de maximisation de la vraisemblance complète prenant en compte une régularisation sur les facteurs latents. Cela permet d'obtenir la loi a posteriori des coefficients de régression du modèle étudié. De plus, la philosophie d'un algorithme EM est différente de celle d'un échantillonneur de Gibbs puisque l'on cherche à maximiser la vraisemblance complète dans un cas, tandis que l'on échantillonne par simulations dans l'autre cas. De plus, nous avons obtenu les estimateurs de la vraisemblance marginale de U et σ^2. Dans l'échantillonnage de Gibbs, la variance du résidu est mise à jour à chaque itération. Pour la factorisation de matrice bayésienne de Salakhutdinov and Mnih (2008), cette valeur est fixée. Une idée, pour simplifier l'échantillonneur de Gibbs de LFMM, serait d'utiliser l'estimateur de la vraisemblance marginale de σ^2 pour fixer la variance résiduelle de l'échantillonneur de Gibbs. Il se peut, toutefois, que cet estimateur de maximum de vraisemblance marginale soit différent de celui estimé par échantillonnage de Gibbs. Enfin, une difficulté majeure de cette approche probabiliste est l'estimation de la variance a priori des coefficients de régression, D_B.

Le problème MAP est, en fait, une généralisation du problème Ridge, analysé prédécemment. Nous avons proposé un algorithme itératif, de type moindre carrés alternés, pour déterminer ces estimateurs. Cet algorithme itératif ne nous garantit pas l'obtention d'un maximun global. Toutefois, on peut utiliser ces estimateurs comme initialisation de l'échantillonneur de Gibbs de LFMM.

L'algorithme Variational Bayes est un algorithme itératif qui approche la distribution a posteriori en faisant certaines hypothèses sur cette dernière. Cet algorithme est intéressant car il est peu coûteux en calcul. Cependant, certaines études et nos premiers tests nous ont montré que cet algorithme a tendance à sous estimer la variance des estimateurs (Jaakkola and Jordan, 2000). Ceci est problématique pour LFMM car on cherche à effectuer un test.

En conclusion, nous avons proposé un ensemble d'estimateurs et d'algorithmes alternatifs à l'échantillonnage de Gibbs. La difficulté principale réside dans l'estimation des variances des coefficients de régression afin de pouvoir effectuer un test. Cependant, ces estimateurs présentent des avantages et des limites qu'il serait intéressant d'investiguer en pratique grâce à des études de simulations, dans la continuité de cette thèse.

Notes bibliographiques

Les articles suivants ont été publiés au cours de la thèse.

Frichot E, François O. 2014. LEA : an R package for Landscape and Ecological Association studies. Manuscrit en cours de rédaction.

De Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE. 2014. Genome scan methods against more complex models : when and how much should we trust them ? Mol Ecol. 23 : 2006–2019.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196 :973–983.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 30 : 1687–1699.

Frichot E, Schoville SD, Bouchard G, François O. 2012. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. Front Genet. 3 :254.

Jay F, Blum MGB, Frichot E, François O. 2011. Modèles à variables latentes en génétique des populations. JSFdS. 152(3).

Bibliographie

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467 :1061–1073.

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491 :56–65.

Abney M, Ober C and McPeek MS. 2002. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees : fasting serum-insulin level in the hutterites. Am J Hum Genet. 70 :920–934.

Ahmed N, Natarajan T and Rao KR. 1974. Discrete cosine transfom. IEEE T Comput. C-23 :90–93.

Akey JM. 2009. Constructing genomic maps of positive selection in humans : where do we go from here ? Genome Res. 19 :711–722.

Alexander DH and Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12 :246.

Alexander DH, Novembre J and Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19 :1655–1664.

Anderson JC and Gerbing DW. 1984. Statistical inference in factor analysis. Psychometrika 49 :155–173.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408 :796.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465 :627–631.

Aulchenko YS, Ripke S, Isaacs A and Van Duijn CM. 2007. GenABEL : an R library for genome-wide association analysis. Bioinformatics 23 :1294–1296.

Balding DJ and Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96 :3–12.

Barrett RD and Hoekstra HE. 2011. Molecular spandrels : tests of adaptation at the genetic level. Nat Rev Genet. 12 :767–780.

Beaumont MA and Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 13 :969–980.

Beaumont MA and Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. P Roy Soc B-Bio Sci. 263 :1619–1626.

Belle E and Barbujani G. 2007. Worldwide analysis of multiple microsatellites : language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol. 133 :1137–1146.

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate : a practical and powerful approach to multiple testing. J R Stat Soc B Met. pp. 289–300.

Berry A and Kreitman M. 1993. Molecular analysis of an allozyme cline : Alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. Genetics 134 :869–893.

Berry MW, Browne M, Langville AN, Pauca VP and Plemmons RJ. 2007. Algorithms and applications for approximate nonnegative matrix factorization. Comput Stat Data An. 52 :155–173.

Blair LM, Granka JM and Feldman MW. 2014. On the stability of the Bayenv method in assessing human SNP-environment associations. Hum Genomics 8 :1.

Borcard D and Legendre P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol Model. 153 :51–68.

Borcard D, Legendre P, Avois-Jacquet C and Tuomisto H. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85 :1826–1832.

Brunet JP, Tamayo P, Golub TR and Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. P Natl Acad Sci. 101 :4164–4169.

Cardon LR and Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet. 2 :91–99.

Carl G and Kühn I. 2007. Analyzing spatial autocorrelation in species distributions using gaussian and logit models. Ecol Model. 207 :159–170.

Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q and West M. 2008. High-dimensional sparse factor modeling : applications in gene expression genomics. J Am Stat Assoc. 103 :1438–1456.

Cavalli-Sforza LL. 2005. The human genome diversity project : past, present and future. Nat Rev Genet. 6 :333–340.

Cavalli-Sforza LL and Bodmer WF. 1971. The genetics of human populations. New York (NY) : Courier Dover Publications.

Chen C, Durand E, Forbes F and François O. 2007. Bayesian clustering algorithms ascertaining spatial population structure : a new computer program and a comparison study. Mol Ecol Notes 7 :747–756.

Chen C, Forbes F and François O. 2006. Fastruct : model-based clustering made faster. Mol Ecol Notes 6 :980–983.

Chung NC and Storey JD. 2013. Statistical significance of variables driving systematic variation. arXiv preprint arXiv :1308.6013 .

Coop G, Witonsky D, Di Rienzo A and Pritchard JK. 2010. Using environmental correlations to identify loci underlying local adaptation. Genetics 185 :1411–1423.

Cressie N A C 1993. Statistics for spatial data. Revised ed. New York (NY) : Wiley.

Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London : John Murray.

De Villemereuil P, Frichot E, Bazin E, François O and Gaggiotti OE. 2014. Genome scan methods against more complex models : when and how much should we trust them ? Mol Ecol. 23 :2006–2019.

Devlin B and Roeder K. 1999. Genomic control for association studies. Biometrics 55 :997–1004.

Diaconis P, Goel S and Holmes S. 2008. Horseshoes in multidimensional scaling and local kernel methods. Ann Appl Stat. 2 :777–807.

Ding C, Li T and Peng W. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Comput Stat Data An 52 :3913–3927.

Dray S, Legendre P and Peresneto P. 2006. Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Model. 196 :483–493.

Duforet-Frebourg N, Bazin E and Blum MGB. 2014. Genome scans for detecting footprints of local adaptation using a bayesian factor model. Mol Biol Evol. p. msu182.

Duforet-Frebourg N and Blum MGB 2014. Nonstationary patterns of isolation-by-distance : inferring measures of local genetic differentiation with bayesian kriging. Evolution 68 :1110–1123.

Dunn OJ. 1961. Multiple comparisons among means. J Am Stat Assoc. 56 :52–64.

Durand E, Jay F, Gaggiotti OE and François O 2009. Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol. 26 :1963–1973.

Eastment HT and Krzanowski WJ. 1982. Cross-validatory choice of the number of components from a principal component analysis. Technometrics 24 :73–77.

Eckart C and Young G. 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 :211–218.

Eckert AJ, Bower AD, Gonzàlez-Martìnez SC, Wegrzyn JL, Coop G and Neale DB. 2010. Back to nature : ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Mol Ecol. 19 :3789–3805.

Eden E, Navon R, Steinfeld I, Lipson D and Yakhini Z. 2009. GOrilla : a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10 :48.

Endler JA. 1977. Geographic variation, speciation, and clines. Princeton (NJ) : Princeton University Press.

Engelhardt BE and Stephens M 2010. Analysis of population structure : a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6 :12.

Evanno G, Regnaut S and Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. Mol Ecol. 14 :2611–2620.

Falush D, Stephens M and Pritchard JK. 2003. Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. Genetics 164 :1567–1587.

Falush D, Stephens M and Pritchard JK. 2007. Inference of population structure using multilocus genotype data : dominant markers and null alleles. Mol Ecol Notes 7 :574–578.

Foll M and Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers : a bayesian perspective. Genetics 180 :977–993.

François O, Currat M, Ray N, Han E, Excoffier L and Novembre J. 2010. Principal component analysis under population genetic models of range expansion and admixture. Mol Biol Evol. 27 :1257–68.

François O and Durand E. 2010. Spatially explicit bayesian clustering models in population genetics. Mol Ecol Resour. 10 :773–784.

François O, Blum Michael GB, Jakobsson M and Rosenberg NA. 2008. Demographic history of european populations of *Arabidopsis thaliana*. PLoS Genet. 4 :e1000075.

Frichot E, Schoville SD, Bouchard G and François O. 2012. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. Front Genet. 3 :254.

Frichot E, Schoville SD, Bouchard G and François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 30 :1687–1699.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L and Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 7 :e1002355.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Chang LY, Huang W, Liu B, Shen Y et al. 2003. The international hapmap project. Nature 426 :789–796.

Günther T and Coop G. 2013. Robust identification of local adaptation from allele frequencies. Genetics 195 :205–220.

Haldane JBS. 1948. The theory of a cline. J Genet. 48 :277–284.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G and Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 7 :16.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G and Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 4 :13.

Harmon LJ and Glor RE. 2010. Poor statistical performance of the Mantel test in phylogenetic comparative analyses. Evolution 64 :2173–2178.

Hijmans RJ, Cameron SE, Parra JL, Jones PG and Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 25 :1965–1978.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS and Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. P Natl Acad Sci USA. 106 :9362–9367.

Hoerl AE and Kennard RW. 1970. Ridge regression : Biased estimation for nonorthogonal problems. Technometrics 12 :55–67.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA and Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. PLoS genetics 6 :e1000862.

Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V and Sullivan M. 2012. PhosphoSitePlus : a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 40(D1) :D261–D270.

Hubisz MJ, Falush D, Stephens M and Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 9 :1322–1332.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18 :337–338.

Jaakkola TS and Jordan MI. 2000. Bayesian parameter estimation via variational methods. Stat Comp. 10 :25–37.

Jakobsson M and Rosenberg NA. 2007. CLUMPP : a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23 :1801–1806.

Jay F, Manel S, Alvarez N, Durand EY, Thuiller W, Holderegger R, Taberlet P and François O. 2012. Forecasting changes in population genetic structure of alpine plants in response to global warming. Mol Ecol. 21 :2354–68.

Johnson VE. 2013. Revised standards for statistical evidence. P Natl A Sci. 110 :19313–19317.

Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat. 29 :295–327.

Jolliffe I T 1986. Principal Component Analysis. New York (NY) : Springer Verlag.

Jombart T, Devillard S, Dufour AB and Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101 :92–103.

Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G and Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection : Towards a landscape genomics approach to adaptation. Mol Ecol. 16 :3955–3969.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E et al. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 42 :348–354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ and Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics 178 :1709–1723.

Kelley JL, Madeoy J, Calhoun JC, Swanson W and Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16 :980–989.

Kim H and Park H. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23 :1495–1502.

Kim J and Park H. 2011. Fast nonnegative matrix factorization : an active-set-like method and comparisons. SIAM J Sci Comput. 33 :3261–3281.

Kimura M and Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49 :561–576.

Koren Y, Bell R and Volinsky C. 2009. Matrix factorization techniques for recommender systems. Computer 42 :30–37.

Kort H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O and Mergeay J. 2014. Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across europe in the tree species *alnus glutinosa*. Mol Ecol. doi : 10.1111/mec.12813.

Kruskal JB. 1978. Multidimensional scaling. Sage University Paper series on Quantitative Application in the Social Sciences.

Lawson DJ and Falush D. 2011. Population identification using genetic data. Ann Rev Genom Hum G. 13 :337–361.

Lawson DJ, Hellenthal G, Myers S and Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8 :e1002453.

Lee DD and Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 :788–791.

Lee S, Zou F and Wright FA. 2010. Convergence and prediction of principal component scores in high-dimensional settings. Ann Stat. 38 :3605.

Legendre P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J Stat Comput Sim. 67 :37–73.

Legendre P and Gallagher E. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129 :271–280.

Legendre P and Legendre L. 2012. Numerical ecology. 3rd English ed. Amsterdam : Elsevier.

Lenormand T. 2002. Gene flow and the limits to natural selection. Trends Ecol Evol. 17 :183–189.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL and Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319 :1100–1104.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI and Heckerman D. 2011. Fast linear mixed models for genome-wide association studies. Nat Methods 8 :833–835.

Luikart G, England PR, Tallmon D, Jordan S and Taberlet P. 2003. The power and promise of population genomics : from genotyping to genome typing. Nat Rev Genet. 4 :981–994.

Malécot G 1948. Les Mathématiques de l'Hérédité. Paris : Masson.

Manel S, Joost S, Epperson BK, Holderegger R, Storfer A, Rosenberg MS, Scribner KT, Bonin A and Fortin MJ. 2010. Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Mol Ecol. 19 :3760–3772.

Manel S, Schwartz MK, Luikart G and Taberlet P. 2003. Landscape genetics : combining landscape ecology and population genetics. Trends Eco Evol. 18 :189–197.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27 :209–220.

Marchini J, Cardon LR, Phillips MS and Donnelly P. 2004. The effects of human population structure on large genetic association studies. Nat Genet. 36 :512–517.

Mardia KV, Kent JT and Bibby JM. 1979. Multivariate Analysis. London : Academic Press.

McCullagh P and Nelder JA. 1989. Generalized linear models. London : Chapman and Hall.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5 :10.

Meirmans PG. 2012. The trouble with isolation by distance. Mol Ecol. 21 :2839–46.

Nakajima S, Sugiyama M and Babacan SD. 2011. Global solution of fully-observed variational bayesian matrix factorization is column-wise independent. *In* Adv Neur In, pp. 208–216.

Nakajima S, Sugiyama M, Babacan SD and Tomioka R. 2013. Global analytic solution of fully-observed variational bayesian matrix factorization. J Mach Learn Res. 14 :1–37.

Nei M. 1972. Genetic distance between populations. Am Nat. 106 :283–292.

Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K and Donnelly P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. J R Stat Soc B 64 :695–715.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39 :197–218.

Novembre J and Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet. 10 :745–755.

Novembre J and Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40 :646–649.

Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 74 :765–769.

Parry RM and Wang MD. 2013. A fast least-squares algorithm for population inference. BMC bioinformatics 14 :28.

Patterson N, Price AL and Reich D 2006. Population structure and eigenanalysis. PLoS Genet. 2 :20.

Patterson Nick J., Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T and Reich D. 2012. Ancient admixture in human history. Genetics 192 :1065–1093.

Pavlidis P, Jensen JD, Stephan W and Stamatakis A. 2012. A critical assessment of storytelling : gene ontology categories and the importance of validating genomic scans. Mol Biol Evol. 29 :3237–3248.

Pearson K. 1901. Liii. on lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci. 2 :559–572.

Poncet BN, Herrman D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S and Manel S. 2010. Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. Mol Ecol. 19 :2896–2907.

Powell GT, Yang H, Tyler-Smith C and Xue Y. 2007. The population history of the Xibe in northern China : a comparison of autosomal, mtDNA and Y-chromosomal analyses of migration and gene flow. Forensic Sci Int-Gen. 1 :115–119.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38 :904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D and Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5 :e1000519.

Price AL, Zaitlen NA, Reich D and Patterson N. 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 11 :459–463.

Pritchard JK, Pickrell JK and Coop G. 2010. The genetics of human adaptation : hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 20 :208–215.

Pritchard JK, Stephens M and Donnelly P. 2000a. Inference of population structure using multilocus genotype data. Genetics 155 :945–959.

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P. 2000b. Association mapping in structured populations. Am J Hum Genet. 67 :170–181.

Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V and Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol. 15 :1022–1027.

R Development Core Team. 2012. R : a language and environment for statistical computing. Vienna : R Foundation for Statistical Computing.

Raiko T, Ilin A and Karhunen J. 2007. Principal component analysis for large scale problems with lots of missing values, pp. 691–698. *In* Lect Notes Artif Int. Springer.

Raj A, Stephens M and Pritchard JK. 2014. Variational inference of population structure in large snp datasets. Genetics 114 :164350.

Roberts DF and Hiorns RW 1965. Methods of analysis of the genetic composition of a hybrid population. Hum Biol. 37 :38–43.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW. 2002. Genetic structure of human populations. Science 298 :2381–2385.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145 :1219–1228.

Saccone SF, Quan J, Mehta G, Bolze R, Thomas P, Deelman E, Tischfield JA and Rice JP. 2011. New tools and methods for direct programmatic access to the dbSNP relational database. Nucleic Acids Res. 39 :901–907.

Salakhutdinov R and Mnih A. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. ICML 2008 25 :880–887.

Sammel MD and Ryan LM. 1996. Latent variable models with fixed effects. Biometrics 52 :650–663.

Sammon JW. 1969. A nonlinear mapping for data structure analysis. IEEE Trans. Comp. 18 :401–409.

Sánchez BN, Budtz-Jørgensen E, Ryan LM and Hu H. 2005. Structural equation models : a review with applications to environmental epidemiology. J Am Stat Assoc. 100 :1443–1455.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C and Manel S. 2012. Adaptive genetic variation on the landscape : Methods and cases. Annu Rev Ecol Syst. 43 :23–43.

Seeger M and Bouchard G. 2012. Fast variational bayesian inference for non-conjugate matrix factorization models. *In* Aistats.

Segelbacher G, Cushman SA, Epperson BK, Fortin MJ, Francois O, Hardy OJ, Holderegger R, Taberlet P, Waits LP and Manel S. 2010. Applications of landscape genetics in conservation biology : concepts and challenges. Conserv Genet. 11 :375–385.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB et al. 2010. Genetic evidence for high-altitude adaptation in tibet. Science 329 :72–75.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139 :457–462.

Slatkin M and Arter HE. 1991. Spatial autocorrelation methods in population genetics. Am Nat. 138 :499–517.

Smouse PE, Long JC and Sokal RR. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zoo. 35 :627–632.

Sokal RR and Oden NL. 1978. Spatial autocorrelation in biology : 1. methodology. Biol J Lin Soc. 10 :199–228.

Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A. 2002. Bayesian measures of model complexity and fit. J R Stat Soc B. 64 :583–639.

Srebro N, Rennie J and Jaakkola TS. 2004. Maximum-margin matrix factorization. *In* Advances in neural information processing systems, pp. 1329–1336.

Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc B 64 :479–498.

Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol. 14 :671–688.

Storz JF and Wheat CW. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. Evolution 64 :2489–2509.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S et al. 2014. High performance computation of landscape genomic models integrating local indices of spatial association. arXiv preprint arXiv :1405.7658 .

Tang H, Peng J, Wang P and Risch NJ. 2005. Estimation of individual admixture : analytical and study design considerations. Genet Epidemiol. 28 :289–301.

Thibert-Plante X and Hendry AP. 2010. When can ecological speciation be detected with neutral loci ? Mol Ecol. 19 :2301–2314.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc B pp. 267–288.

Tikhonov AN. 1943. On the stability of inverse problems. *In* Dokl. Akad. Nauk SSSR, volume 39, pp. 195–198.

Tipping ME and Bishop CM. 1999. Probabilistic principal component analysis. J R Stat Soc B. 61 :611–622.

Tracy CA and Widom H. 1994. Level spacing distributions and the bessel kernel. Commun Math Phys. 161 :289–309.

Weigel D and Mott R. 2009. The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. 10 :107.

West M. 2003. Bayesian factor regression models in the "large p, small n" paradigm. Bayes Stat. 7 :723–732.

Williams GC. 1966. Adaptation and Natural Selection, volume 1996. Princeton (NJ) : Princeton University Press.

Wold S. 1978. Cross-Validatory estimation of the number of components in factor and principal components models. Technometrics 20 :397–405.

Woodard DB, Love TMT, Thurston SW, Ruppert D, Sathyanarayana S and Swan SH 2013. Latent factor regression models for grouped outcomes. Biometrics 69 :785–794.

Wright S. 1943. Isolation by distance. Genetics 28 :114–138.

Wu B, Liu N and Zhao H. 2006. Psmix : an r package for population structure inference via maximum likelihood method. BMC bioinformatics 7 :317.

Young JH, Chang YC, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY and Chakravarti A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. Plos Genet. 1 :9.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38 :203–208.

Zhang F, Su B, Zhang Y and Jin L. 2007. Genetic studies of human diversity in East Asia. Philos T Roy Soc B. 362 :987–996.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 42 :355–360.

Zhou X, Carbonetto P and Stephens M. 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 9 :e1003264.

Zhou X and Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 44 :821–4.

Zueva KJ, Lumme J, Veselov AE, Kent MP, Lien Sigbjørn and Primmer CR. 2014. Footprints of directional selection in wild atlantic salmon populations : evidence for parasite-driven evolution ? PloS one 9 :e91672.

Titre : Modèles a facteurs latents pour les études d'association écologique en génétique des populations.

Résumé : Nous introduisons un ensemble de modèles à facteurs latents dédié à la génomique du paysage et aux tests d'associations écologiques. Cela comprend des méthodes statistiques pour corriger des effets d'autocorrélation spatiale sur les cartes de composantes principales en génétique des populations (spFA), des méthodes pour estimer rapidement et efficacement les coefficients de métissage individuel à partir de matrices de génotypes de grande taille et évaluer le nombre de populations ancestrales (sNMF) et des méthodes pour identifier les polymorphismes génétiques qui montrent de fortes corrélations avec des gradients environnementaux ou avec des variables utilisées comme des indicateurs pour des pressions écologiques (LFMM). Nous avons aussi développé un ensemble de logiciels libres associés à ces méthodes, basés sur des programmes optimisés en C qui peuvent passer à l'échelle avec la dimension de très grand jeu de données, afin d'effectuer des analyses de structures de population et des cribles génomiques pour l'adaptation locale.

Mots-clés : modèles à facteurs latents, adaptation locale, structure génétique des populations, séquencage haut-debit, statistiques bayésiennes, apprentissage.

Title : Latent factor models for ecological association studies in population genetics.

Summary : We introduce a set of latent factor models dedicated to landscape genomics and ecological association tests. It includes statistical methods for correcting principal component maps for effects of spatial autocorrelation (spFA) ; methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (sNMF) ; and methods for identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures (LFMM). We also developed a set of open source softwares associated with the methods, based on optimized C programs that can scale with the dimension of very large data sets, to run analyses of population structure and genome scans for local adaptation.

Keywords : latent factor models, local adaptation, population genetic structure, next generation sequencing, bayesian statistics, machine learning.