RAPPORT DE STAGE DE FIN D'ETUDES

Présenté par **Alia DEHMAN**

Mission effectuée du *1er Février* au *31 Juillet 2012* au :

Laboratoire Statistique et Génome

Sujet de la mission

Régression parcimonieuse structurée pour les études d'association pangénomiques

Directeur de Stage : **Christophe AMBROISE**
Co-Directeur de Stage : **Pierre NEUVIAL**
Conseiller de Stage : **Wojciech PIECZYNSKI**

Remerciements

Je voudrais, tout d'abord, remercier le Professeur Christophe AMBROISE qui m'a accueillie au sein de son équipe et m'a guidée dans mon travail avec beaucoup de patience et de disponibilité. Je remercie également Monsieur Pierre NEUVIAL pour son aide précieuse, ses conseils avisés et pour m'avoir fait confiance tout au long de ce stage.

Je voudrais aussi remercier Monsieur Julien CHIQUET et Monsieur Cyril DALMASSO pour avoir répondu à mes nombreuses questions, aussi bien en biologie qu'en statistique.

Je tiens à exprimer des remerciements particuliers à Monsieur Bernard Prum, Professeur Emérite au Laboratoire Statistique et Génome pour son soutien constant et pour toute l'aide qu'il m'a apportée.

Je remercie mon Professeur à Télécom SudParis, Monsieur Randal DOUC, pour m'avoir orientée vers cette belle discipline des statistiques et je remercie également le Professeur Wojciech PIECZYNSKI pour m'avoir accompagnée et conseillée pendant ce stage.

Enfin, j'exprime mes remerciements à tous les membres du laboratoire pour l'excellente ambiance et l'accueil chaleureux qu'ils m'ont réservé.

Table des matières

Introduction

La séquence de l'ADN contient l'information nécessaire aux êtres vivants pour vivre et se reproduire. C'est pourquoi, l'exploitation et l'analyse de ces données génomiques, bien qu'il soit un travail assez laborieux et de longue haleine, contribue de manière spectaculaire à l'avancement des recherches dans les domaines de la physiologie ou de la pathologie humaine.
Par ailleurs, l'avancée des technologies, a permis de mettre à la disposition de la recherche un grand nombre de données très diversifiées : à côté des séquences chromosomiques, on dispose actuellement de séquences protéiques, de donnés phénotypiques, etc.
Des informations pertinentes ne peuvent être tirées de cet amas de données que si l'on dispose :

- d'outils puissants pour stocker ces données
- de méthodes scientifiques performantes pour les analyser

Ce deuxième point nécessite justement la collaboration, au sein des équipes de recherche, entre informaticiens, statisticiens et biologistes pour pouvoir mettre en œuvre des algorithmes efficaces de point de vue mathématique, et qui soient fidèles à la réalité biologique du problème.

L'utilisation de ce grand nombre de données génomiques dans l'analyse de données d'association à grande échelle représente une nouvelle thématique de recherche pour le Laboratoire Statistique et Génome. Mon stage de fin d'études s'inscrit dans cette thématique.

Chapitre 1

Résumé

1.1 Présentation du sujet

Un grand nombre de pathologies ont une composante génétique. Dans le cas le plus simple, le changement d'une lettre dans la séquence d'ADN peut à lui seul être responsable d'une maladie : on parle alors de *maladie monogénique*. De manière moins triviale, les maladies dites *multi-factorielles* ou complexes sont le résultat de composantes multi-géniques et environnementales. C'est le cas de la plupart des cancers, des maladies psychiatriques, auto-immunes et bien d'autres.

Les études épidémiologiques fondées sur la génétique cherchent donc à identifier les loci de susceptibilité et à en quantifier l'influence. Une approche populaire consiste à collecter un échantillon d'individus appelés 'cas' et d'individus non-affectés appelés 'témoins' et de déterminer les positions dans le génome pour lesquelles le texte génétique diffère significativement entre les cas et les témoins. On parle alors d'*étude d'association cas-témoins*. On travaille typiquement à partir d'un jeu de loci appelés *marqueurs génétiques*, dont la position sur le génome est connue. Les marqueurs les plus couramment utilisés sont les *Single Nucleotide Polymorphisms* (SNP). Définis comme des positions sur le chromosome où le texte génétique varie d'une seule base d'un individu à un autre, ils sont nombreux : plusieurs millions sur les 3 milliards de lettres (ou bases) qui constituent le génome humain. La diminution à la fois du coût et du temps de génotypage des SNP a contribué récemment au lancement d'études génétiques d'association à grande échelle permettant d'explorer une part conséquente des polymorphismes génétiques pouvant être impliqués dans les mécanismes biologiques à l'origine des maladies. L'analyse de telles données soulève de nombreuses questions méthodologiques.

En particulier, les phénomènes de recombinaison induisent une structure de dépendance entre SNP (appelée *déséquilibre de liaison*) qui est rarement prise en compte dans les études d'association. Mon stage a donc visé à prendre en compte cette dépendance dans le cadre de modèles particuliers de régression logistique parcimonieuse, avec l'espoir d'obtenir des modèles à la fois plus performants, et surtout plus faciles à interpréter du point de vue biologique. J'ai donc commencé par comprendre les modèles existants, avant de proposer des alternatives originales qui prennent en compte ce phénomène de déséquilibre de liaison pour enfin comparer les propositions sur des simulations réalistes.

Quelques définitions des termes les plus importants seront présentées, plus en détails, plus loin dans le rapport.

1.2 Outils utilisés

Les algorithmes pour tester les différentes méthodes ont été développés avec le langage de programmation R. C'est un langage particulièrement adapté pour traiter les problèmes mathématiques et statistiques.
Il s'agit d'un langage open source qui met à disposition de l'utilisateur un certain nombre de fonctions prédéfinies, appelées *packages*, notamment dans le domaine de la biostatistique.
Cependant, ces fonctions standard fournies par R ne sont parfois pas adaptées aux problématiques spécifiques de chacun. C'est pourquoi, il est nécessaire de bien se documenter sur les packages avant de les utiliser ou bien coder les fonctions dont on a besoin.

Pour la rédaction de ce rapport, j'ai utilisé LaTeX. Il s'agit d'un langage de description de document permettant une mise en page très professionnelle. Les documents créés en LaTeX étant très clairs, lisibles et optimisés pour l'affichage de formules mathématiques, ce langage est, de nos jours, un outil de choix pour les scientifiques pour la rédaction des rapports et des articles scientifiques.

Chapitre 2

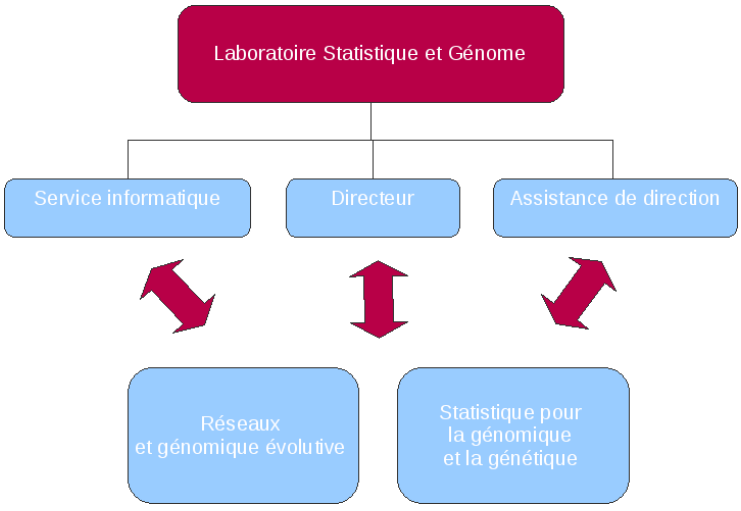
Présentation du laboratoire

2.1 Présentation générale et activités

Le Laboratoire Statistique et Génome est un laboratoire de recherche affilié au CNRS, à l'Université d'Evry Val d'Essonne et à l'INRA. Y évoluent deux lignes de recherches principales : *Réseaux et Génomique évolutive* et *Statistique pour la génomique et la génétique*.
La recherche au sein du laboratoire a pour objectif de développer des méthodes originales pour l'analyse de données biologiques, issues majoritairement de la biologie moléculaire.

2.2 Organigramme

Ci-dessous l'organigramme du laboratoire illustrant les deux lignes de recherche qui le constituent ainsi que les personnes qui y travaillent.



Le laboratoire reçoit également des doctorants et des stagiaires. Il accueille entre 5 et 8 stagiaires chaque année et, pendant la période de mon stage, y évoluaient 7 doctorants.

2.3 Rayonnement du laboratoire

L'existence du Laboratoire Statistique et Génome est due à la présence de la *Génopole*, un campus de recherche d'excellence en génomique. Le Laboratoire Statistique et Génome est reconnu par la Génopole comme le laboratoire de statistique de référence, essentiel pour le développement de méthodes statistiques pertinentes dans le domaine de la biologie.

Chapitre 3

Les motifs du choix du stage

J'ai choisi d'effectuer mon stage de fin d'études au Laboratoire Statistique et Génome parce que je voulais continuer dans le domaine de la recherche en mathématiques appliquées, un domaine que j'ai déjà découvert à travers mon premier stage de découverte que j'ai effectué au Centre Basque de Recherche en Mathématiques Appliquées à Bilbao et dans lequel j'ai choisi de continuer d'évoluer en suivant la voie d'approfondissement "Modélisation Statistique et Applications" à Télécom SudParis.

Plus encore, ce laboratoire m'a proposé la possibilité de poursuivre mon travail de stage en réalisant une thèse dans le domaine de la biostatistique. Le sujet de thèse qui prolonge le sujet de stage est en parfait lien avec mon projet professionnel puisqu'il me permettra notamment de développer des méthodes statistiques pointues pour les études d'association pangénomiques. Ce projet me donnera aussi l'occasion d'implémenter ces méthodes de manière optimale et d'interpréter, d'un point de vue biologique, les résultats obtenus. Le caractère pluridisciplinaire de ce projet liant à la fois les statistiques, l'informatique et la biologie m'intéresse tout particulièrement.

Aujourd'hui, je souhaite effectuer des travaux de recherche au sein d'un institut de recherche publique, c'est également pour cela que j'ai choisi le Laboratoire Statistique et Génome pour réaliser mon stage de fin d'études. La réalisation d'une thèse s'inscrit dans la logique de ma formation, en même temps que dans la perspective de mes ambitions professionnelles. Je souhaite effectivement effectuer ma carrière dans l'enseignement et la recherche.

Chapitre 4

Contexte et objectifs du stage

Dans ce chapitre, nous allons d'abord présenter quelques définitions biologiques liées à la génomique. Puis, nous présenterons les notions de statistiques sur lesquelles se sont basés les travaux durant le stage pour pouvoir ensuite exposer la modélisation statistique de la problématique traitée. Nous finirons par détailler les objectifs du stage.

4.1 Quelques notions de biologie

4.1.1 Polymorphisme d'un seul nucléotide

On appelle *locus* une position du génome et *allèle* une version donnée du texte génétique.

Le *polymorphisme génétique* (du grec "poly" plusieurs et "morphe" forme) est défini comme la coexistence de plusieurs allèles pour un gène ou locus donnés, dans une population.
Dans l'espèce humaine, chaque individu possède 22 paires de chromosomes homologues. Ainsi, en un locus donné, il y a une combinaison donnée de deux allèles appelée *génotype*.

Les *polymorphismes d'un seul nucléotide*, en anglais *Single Nucleotide Polymorphisms* (SNP) sont les polymorphismes les plus répandus du génome humain. Il s'agit d'une variation d'une seule paire de bases du génome entre individus d'une même espèce. Par exemple, deux séquences de fragments d'ADN de différents individus, AAGC**C**TA et AAGC**T**TA diffèrent en un seul nucléotide. Dans ce cas, on parle alors d'un *polymorphisme biallélique*, c'est-à-dire de deux allèles : C et T. Presque tous les SNP sont biallélliques.

Ces polymorphismes nucléotidiques sont répartis sur l'ensemble du génome, toutes les 500 paires de bases en moyenne. Et à l'heure actuelle, plus de 5 millions de SNP ont été caractérisés.

4.1.2 Puces à ADN

Une puce à ADN, alias une *biopuce*, est un petit outil d'analyse et de diagnostic d'environ 1 cm^2. Elle se présente sous la forme d'un support en verre ou en silicium sur lequel sont fixés des milliers de protéines ou de fragments d'ADN (ou d'ARN).
Cette biotechnologie récente permet de visualiser les séquences de bases de l'ADN dans une cellule d'un tissu donné (foie, intestin...), à un moment donné (embryon, adulte...) et dans un état donné (malade, sain...). Ces molécules d'ADN fixées sont appelées *des sondes*. Des milliers de sondes peuvent être fixées sur une même puce. Notons aussi qu'il peut exister plusieurs sondes pour un même gène.
Les données issues des biopuces posent de nouveaux défis aux statisiciens. Ce sont des tableaux classiques individus/variables, mais leur particularité est de décrire très peu d'individus par beaucoup de variables. L'exploitation efficace de ces données requiert donc à la fois une connaissance des processus de fabrication des puces ainsi que des connaissances en statistiques.

4.1.3 Le déséquilibre gamétique

L'objectif des études d'associations demeure celui de trouver des associations entre un phénotype donné et les génotypes de SNP particuliers. Cependant, dans des cas particuliers, à l'issu de l'étude, peuvent ressortir des SNP qui ne sont pas directement liés au phénotype étudié (éventuellement la maladie) mais qui est en réalité en *déséquilibre gamétique* avec un SNP associé à ce phénotype.
Le déséquilibre gamétique ou déséquilibre de liaison, noté LD pour *Linkage Disequilibrium*, est une situation dans laquelle deux allèles, correspondant à deux loci distincts d'un même chromosome, sont plus féquemment associés dans une population que ne le voudrait le hasard. Cette association allélique est favorisée par la proximité physique des deux loci et dépend également des propriétés génomiques de la région.

4.1.4 L'épistasie

La plupart des traits phénotypiques sont contrôlés par plus d'un gène. Le phénomène d'épistasie correspond à l'interaction entre deux ou plusieurs gènes dans le contrôle d'un caractère phénotypique et l'expression d'un de ces gènes peut donc masquer l'expression des autres.

4.2 Outil statistique : la régression linéaire

4.2.1 Le modèle linéaire :

Pour tout i de 1 à N, le modèle linéaire est défini par

$$y_i = \beta_0 + \sum_{j=1}^{p} \mathbf{X}_{ij}\beta_j + \varepsilon_i, \tag{4.1}$$

où $\mathbf{y_i}$ (scalaire) est la réponse à expliquer par le vecteur de variables d'entrées $\mathbf{X_i} = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$. Le scalaire β_0 (appelé biais ou *intercept*) et les $(\beta_j)_{j=1\ldots p}$ sont les paramètres à estimer. Les quantités ε_i sont des scalaires et correspondent aux résidus qu'on

suppose gaussiens, c'est à dire que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
On dispose de N données indépendantes relatives aux variables y et $\mathbf{X}$ (de dimensions $N \times p$) des données génotypiques relatives à chaque individu. Notons $\mathbf{y}$ le vecteur des N observations y_i. Si on rajoute une colonne de 1, au début de la matrice $\mathbf{X}$ alors, le modèle linéaire peut s'écrire comme suit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \tag{4.2}$$

où $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ est le vecteur des paramètres à estimer et ε le vecteur des résidus qu'on suppose indépendants et identiquement distribués.

4.2.2 L'estimateur des moindres carrés

L'*estimateur des moindres carrés ordinaires* est défini comme le vecteur minimisant la somme des carrés résiduels (*RSS* pour *Residual Sum of Squares*) :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} RSS(\beta), \text{ avec } RSS(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2. \tag{4.3}$$

Le théorème suivant donne la solution de ce problème et les grandeurs caractéristiques standards associées à cet estimateur (biais et variance).

Théorème 4.1 *Lorsque* ${}^{\mathbf{t}}\mathbf{X}\mathbf{X}$ *est inversible, la solution* $\hat{\boldsymbol{\beta}}$ *des équations normales est unique et est appelée estimateur de Gauss-Markov :*

$$\hat{\boldsymbol{\beta}} = ({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1}{}^{\mathbf{t}}\mathbf{X}\mathbf{y}.$$

C'est un estimateur sans biais *de* β *et de matrice de covariance*

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1}.$$

4.2.3 Risque quadratique et erreur de prédiction

Définition 4.2 *On appelle* **risque quadratique** *(Mean Squared Error) de l'estimateur* $\hat{\boldsymbol{\beta}}$ *du paramètre* $\boldsymbol{\beta}$ *la grandeur définie par*

$$MSE(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2] = biais^2(\hat{\boldsymbol{\beta}}) + Var(\hat{\boldsymbol{\beta}}).$$

Ainsi, on constate que, le risque quadratique de $\hat{\boldsymbol{\beta}}$ se ramène simplement à sa variance puisqu'il est sans biais.

Définition 4.3 *On appelle* **erreur de prédiction** *(Expected Prediction Error) du modèle de régression* $y = f(X) + \varepsilon$*, la grandeur définie par*

$$EPE(f) = \mathbb{E}[(y - f(X))^2].$$

Dans le cas du modèle linéaire, l'estimateur de Gauss-Markov prévoit pour y le vecteur de valeurs $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Ainsi, nous avons :

$$EPE(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbb{E}[(y - \mathbf{X}\hat{\boldsymbol{\beta}})^2] = (\mathbb{E}[y - \mathbf{X}\hat{\boldsymbol{\beta}}])^2 + Var(\mathbf{y} + \mathbf{X}\hat{\boldsymbol{\beta}}) = \sigma^2 + MSE(\mathbf{X}\hat{\boldsymbol{\beta}}).$$

4.3 Modélisation du problème

Les données découlant d'une étude d'association sur une population donnée sont constituées de deux composants principaux : *le phénotype* ou le trait, qui peut être soit quantitatif soit binaire (malade ou sain en général), *les génotypes* qui représentent les *variables explicatives*.
Par exemple, soit $\mathbf{y}_i$ le phénotype pour l'individu i dans notre échantillon, où $i = 1 \dots N$. N est la taille totale de notre échantillon, c'est-à-dire le nombre d'individus sur lesquels est faite l'étude. x_{ij} le représente le génotype du SNP j pour l'individu $i = 1 \dots p$, où p est le nombre total de SNP à l'étude. On obtient ainsi une matrice $\mathbf{X}$ de taille $N \times p$ de variables génotypiques $\mathbf{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$.

Une fois ces notations posées, on devine donc comment on peut appliquer les théories statistiques citées dans la section 4.2. Des études d'association sur tout le génome, en utilisant la régression linéaire peuvent être menées pour plusieurs raisons :

- pour améliorer la connaissance que l'on a d'un ensemble de phénomènes (notamment des maladies). Par exemple, recherche de relation de causalité entre un phénotype donné et des variables génotypiques.
- dans un but prédictif : il s'agit alors de prédire la variable de sortie (non observable ou dont les valeurs sont difficiles ou coûteuses à obtenir), en fonction des variables d'entrée. Par exemple, prédire si un individu va être malade ou sain pour un phénotype donné, disposant de ses données génotypiques.

Par ailleurs, en général, le génotype au site j pour un individu i, noté x_{ij} est une variable discrète qui prend deux ou plusieurs valeurs. Par exemple, en considérant les SNP bialléliques (a,A), chaque individu sera donc porteur, au niveau d'un SNP d'un des trois génotypes possibles, un des deux génotypes homozygotes (aa et AA) ou le génotype hétérozygote (aA). Par conséquent, x_{ij} sera une variable discrète à trois valeurs. Une autre alternative est de définir x_{ij} comme une variable qui prend seulement deux valeurs discrètes : 0 si le génotype observé est AA et 1 sinon.
Dans la suite du rapport, nous utiliserons les termes *SNP*, *marqueur*, *variable explicative* ou encore *prédicteur*.

4.4 Les objectifs du stage

En choisissant ce stage au Laboratoire Statistique et Génome, mon premier objectif a été de me familiariser avec les outils et les méthodes mathématiques utilisées en biostatistique. En effet, il s'agit d'un domaine d'application des statistiques que je n'ai pas eu l'occasion d'étudier à Télécom SudParis.

En outre, dès le premier entretien, mon encadrant m'a demandé si je voulais continuer avec une thèse si mon stage se passait bien. Je lui ai répondu que c'était justement mon objectif en choisissant ce stage en laboratoire de recherche : commencer une thèse comme continuité de mon stage de fin d'étude, pour pouvoir ensuite enseigner les mathématiques à l'université.

Le sujet que m'a proposé mon encadrant a été un sujet vaste, qui présente une problématique générale posée dans les études d'association pangénomiques. Les objectifs posés par rapport à ce sujet de stage ont donc été doubles : des objectifs à court terme que j'ai atteints pendant mon stage de fin d'études et des objectifs à plus long terme que je traiterai pendant ma thèse.

4.4.1 Problématique globale : le sujet de thèse

Les études d'association pangénomiques visent à identifier des variations génétiques associées à un trait phénotypique donné (malade/sain, charge virale d'une maladie, etc.). Les variations génétiques sont généralement des SNP.

Du point de vue biologique, le contexte est le suivant : on dispose d'un grand nombre de données génomiques (jusqu'à un million de SNP génotypés pour un seul individu) qui ont deux caractéristiques fondamentales : la grande dimension, qui correspond à un très grand nombre de SNP par rapport au nombre d'individus dans notre étude ; et la forte structuration due notamment aux phénomènes d'épistasie et de déséquilibre gamétique. Du point de vue statistique, l'enjeu du projet de thèse est le suivant : les méthodes de régression classiques utilisées par les biologistes sont des méthodes généralement simple-marqueur, c'est-à-dire qui testent l'association au phénotype de chaque SNP individuellement et donc qui ne prennent pas en compte le phénomène d'épistasie. Ces méthodes sont beaucoup moins performantes que l'approche par pénalisation proposée par les statisticiens et encore peu utilisée par les biologistes. Cependant, cette dernière approche n'est pas adaptée au données génomiques réelles à cause de leur forte structuration. L'objectif de la thèse est donc d'adapter les méthodes de régression pénalisée aux données génomiques réelles, en intégrant l'information à priori sur la structure de ces données dans les modèles de régression, et ceci dans le but d'exploiter les propriétés statistiques de la pénalisation dans les applications biologiques.

4.4.2 Objectifs du stage

Les objectifs fixés pour le stage on été les suivants :
- comprendre et maîtriser les notions clés du sujet tels que déséquilibre gamétique, régression linéaire, approche simple-marqueur, approche par pénalisation.
- me familiariser avec les méthodes de régression existantes.
- implémenter ces méthodes existantes afin de comparer leur performance sur des jeux de données simulées.
- faire une étude théorique sur les méthodes existantes de pénalisation dans le but de mettre en évidence les avantages et les limites de chacune, pour pouvoir ensuite proposer une méthode originale qui permette de dépasser ces limites.

Chapitre 5

Etudes menées et résultats

5.1 Les objectifs de la régression linéaire : trois axes d'étude

Un modèle de régression tente d'expliquer au mieux une grandeur Y (la réponse observée) en fonction d'autres grandeurs $\mathbf{x}$ (vecteur de variables explicatives) en quantifiant chacun des aspects déterministe et aléatoire.

En pratique, les applications des méthodes de régression répondent à des objectifs multiples parmi lesquels on peut citer les trois plus importants :

- **estimation des coefficients** de régression $\beta_0, \beta_1, \ldots \beta_p$.
- **prédiction** : lorsque les paramètres du modèle ont été estimés, et en supposant le modèle valide, il est possible de l'utiliser pour *prédire* la valeur que prendra la variable Y pour de nouvelles valeurs des variables explicatives.
- **sélection automatique** des "bonnes" variables explicatives : il est souvent intéressant de déterminer si les résultats de la régression sont dûs au hasard ou s'ils traduisent l'existence d'une relation *significative* entre la variable à expliquer et les variables explicatives.

Durant ce stage, nous nous sommes concentrés sur le troisième objectif, à savoir la conception et le développement de méthodes rapides et pertinentes de sélection automatique de variables.

5.2 Jeux de données utilisés

5.2.1 Données simulées

Pour tester les différentes méthodes de régression linéaire, les données génotypiques ont été simulées comme expliqué au début de la section "analyse de données simulées" de l'article [1].

Notre modèle de simulation diffère un peu de ce qui est proposé dans cet article

puisque ce dernier traite de la régression logistique (c'est-à-dire un modèle avec des sorties binaires). Les simuations ont été réalisées suivant l'équation :

$$y_i = \mathbf{X_i}\boldsymbol{\beta} + {}^{\mathbf{t}}\mathbf{X_i}\boldsymbol{\eta}\mathbf{X_i}. \tag{5.1}$$

Chaque $\mathbf{X_i}$ est dérivé d'une réalisation d'un vecteur normal multivarié $\mathbf{Y_i}$ centré et de matrice de covariance :

$$Cov(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k, \\ \rho & j, k \leq \text{ncorr}, j \neq k, \\ 0 & sinon. \end{cases} \tag{5.2}$$

La variable *ncorr* correspond donc au nombre de variables qu'on simule comme corrélées entre elles. Cette variable peut par exemple prendre la valeur 10. Pour simuler un SNP, on associe à X_{ij} les valeurs -1, 0 ou 1 suivant que $Y_{ij} < -c$, $-c \leq Y_{ij} \leq c$ ou $Y_{ij} > c$. Le seuil c correspond au premier quartile de la distribution normale centrée réduite. $\boldsymbol{\eta}$ est une matrice $p \times p$ symétrique qui représente les interactions entre les marqueurs : si les SNP i et j sont en interaction alors $\boldsymbol{\eta}_{ij}$ est non nul.

D'autre part, pour pouvoir mieux contrôler la difficulté du problème pour les simulations, nous avons introduit la grandeur

$$r^2 = 1 - \frac{||\epsilon||_2^2}{||\mathbf{y}||_2^2}.$$

Ainsi, par définition, la grandeur r^2 prend des valeurs entre 0 et 1. Et le problème est considéré très difficile si r^2 est proche de 0 et il est assez facile si cette grandeur est proche de 1. A partir de l'expression de r^2, on peut déduire l'expression de la variance σ^2 du bruit en fonction de la difficulté du problème :

$$\sigma^2 = \frac{1 - r^2}{r^2}\boldsymbol{\beta}\Sigma\boldsymbol{\beta}.$$

5.2.2 Données réelles

Les données génomiques réelles qui ont été mises à notre disposition sont des données issues d'une étude d'association effectuée par Dalmasso et ses co-auteurs dans l'article [2], pour la recherche de marqueurs biologiques associés à la charge virale de la maladie du sida.
Ces données génomiques consistent en :

- une matrice des génotypes de SNP de $N = 605$ lignes (nombre d'individus) $\times p =$ 307851 colonnes (nombre de SNP).
- un vecteur de longueur N représentant la charge virale de VIH pour chaque individu.
- une matrice des annotations des SNP.

5.3 L'approche simple-marqueur

L'approche simple-marqueur ou l'approche filtre consiste à analyser individuellement l'association de chaque marqueur au trait phénotypique étudié. Statistiquement parlant,

il s'agit d'attribuer à chaque SNP une valeur d'une statistique S, se fixer un seuil en fonction de l'erreur de type I et ensuite, considérer qu'un marqueur est significativement associé au phénotype si sa statistique est supérieure à la valeur seuil.
On aura donc un tableau des résultats de cette forme :

	SNP_1	SNP_2	...	SNP_j	...	SNP_p
S	S_1	S_2	...	S_j	...	S_p
association	non	non	...	oui	...	non

Cette approche a été programmée sur des données réelles et des données simulées.

5.3.1 Démarche et résultats sur données simulées

Le test de significativité d'un marqueur q est un test de l'hypothèse selon laquelle le coefficient β_q est nul. Les hypothèses s'écrivent alors :

$$\begin{cases} H_0 : \beta_q = 0 \\ H_1 : \beta_q \neq 0 \end{cases}$$

Si on appelle S^0_{res} et S^1_{res} la variance résiduelle, respectivement dans le modèle réduit (sous H_0) et dans le modèle complet (sous H_1), alors la statistique

$$F = (N-2)\frac{S^0_{res} - S^1_{res}}{S^1_{res}}$$

suit une loi de Fisher de degrés de liberté 1 et N-2.
Par ailleurs, il existe une fonction R spécifique aux problèmes de régression linéaire et qui calcule, entre autres choses, cette statistique. Il s'agit de la fonction lm.
Nous avons donc programmé l'approche simple-marqueur sur des données simulées de deux manières différentes : en utilisant l'expression explicite de la statistique et en utilisant la fonction lm de R. Voici le code correspondant :

```
#################calcul 'à la main'
p.val <- function(X,phen,N,p){
  p.values <- rep(0,p)
  for (j in 1:p){
   beta.j.hat <- sum(X[,j]*phen)/sum(X[,j]^2)
   phen.hat <- beta.j.hat*X[,j]
##calcul des residus
   S.res.H1 <- sum((phen-phen.hat)^2)
   S.res.H0 <- sum(phen^2)
##calcul de la statistique
   F <- ((S.res.H0-S.res.H1)/S.res.H1)*(N-2)
   p.value <- 1-pf(F,1,N-2)
   p.values[j] <- p.value
  }
  return(p.values)
}
p.valeur <- p.val(X,phen,N,p)[50]
```

Et voici le résultat obtenu pour le calcul de la p-valeur du marqueur numéro 50 :

```
print(summary (lm(phen ~ X[,50] - 1)))
[1] 0.3233538

Call:
lm(formula = phen ~ X[, 50] - 1)
Residuals:
    Min      1Q  Median      3Q     Max
-10.198  -1.719   1.183   3.535  14.037
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
X[, 50]  -0.2483     0.2510   -0.99    0.323
Residual standard error: 3.952 on 499 degrees of freedom
Multiple R-squared: 0.001959,Adjusted R-squared: -4.154e-05
F-statistic: 0.9792 on 1 and 499 DF,  p-value: 0.3229
```

On voit donc que les deux implémentations donnent la même valeur à 10^{-3} près. Dans cette simulation, nous avons pris $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$ c'est-à-dire que les cinq premiers marqueurs sont ceux associés au phénotype. Les cinq premières p-valeurs sont de l'ordre de 10^{-15} et les 95 autres sont supérieures à 0.05. Ainsi, on arrive bien à retrouver les "vrais" SNP associés au phénotype en fixant le seuil d'erreur de type I à 5%.

5.3.2 Résultats sur les données réelles

En testant l'approche simple-marqueur sur les données génomiques réelles, nous avons essayé de retrouver la figure des p-valeurs des SNP du chromosome 6. Les résultats de l'article montrent que les SNP les plus significatifs se trouvent sur ce chromosome. En utilisant la fonction *lm*, voici la figure que nous obtenons :

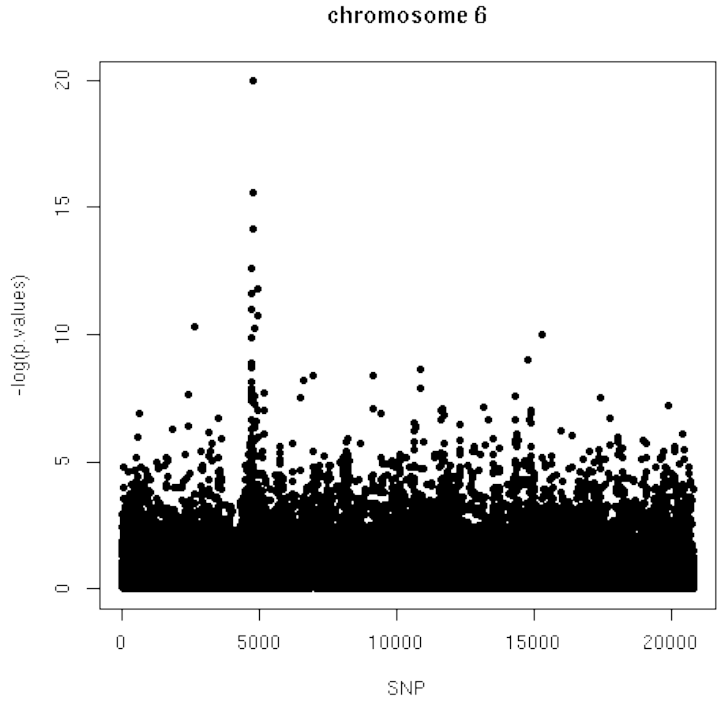


FIGURE 5.1 – Régression univariée des SNP du chromosome 6

et voici la figure reprise de l'article [2] :

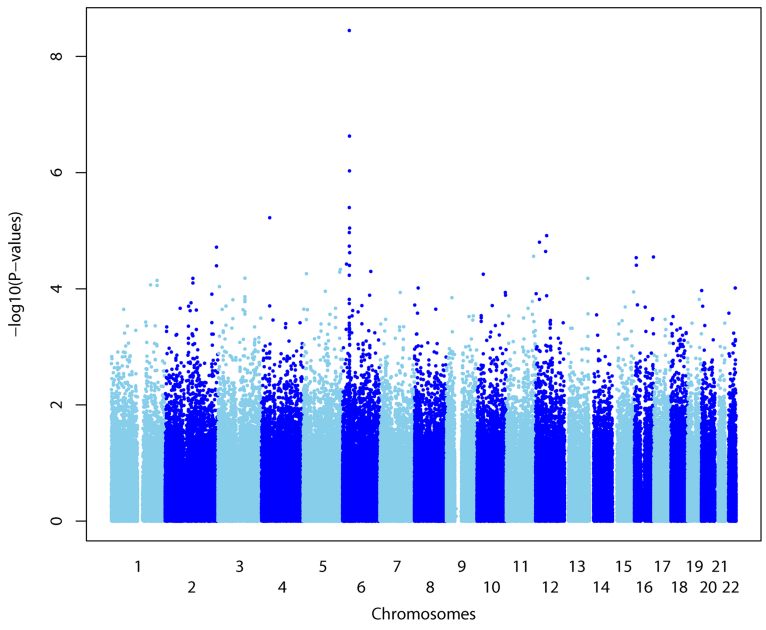


FIGURE 5.2 – Régression univariée des SNP des chromosomes de 1 à 22

Dans les deux figures, en abscisse, chaque point correspond à un marqueur et en ordonnée sont représentés les $-log_{10}$(p-valeurs). Ainsi, les grandes valeurs correspondent à des SNP significativement associés au phénotype.

On constate que la figure 5.1 et la zone correspondant au chromosome 6 de la figure 5.2 ont les mêmes tendances avec un gros pic près du SNP 5000. Cependant, on remarque une différence d'ordre de grandeur des p-valeurs de la figure 5.1 où les plus grandes valeurs sont entre 10 et 20 et celles de l'article [2] où les plus grandes valeurs ne dépassent pas 8. Ce phénomène est du à la structure de la population étudiée. En effet, les études d'association reposent sur une hypothèse très importante, qui est l'hypothèse d'homogénéité de la population à étudier ; une hypothèse qui peut facilement être violée et donc conduire à des erreurs de types I et II. La violation de cette hypothèse conduit à une surestimation de la statistique par un facteur λ qui dépend de l'effet de stratification. Cet effet a été corrigé dans l'article [2]comme une étape de pré-traitement avant de calculer les scores d'association. Et c'est l'absence de cette correction dans mon analyse qui explique cet écart entre les p-valeurs.

5.4 L'approche multivariée

L'utilisation des méthodes de régression multivariée est motivée par le contexte biologique de nos études d'association. En effet, les traits phénotypiques sont généralement contrôlés par plus d'un gène (le phénomène d'épistasie) et donc peuvent être associés à plus d'un SNP. De plus, chaque SNP est associé à d'autres SNP par le phénomène de déséquilibre gamétique, et éventuellement au SNP réellement associé au phénotype étudié.
Dans ce contexte, l'approche simple-marqueur, en étudiant l'effet marginal de chaque marqueur individuellement, semble atteindre ses limites, vu la forte structuration des données génomiques.

Afin de contourner cette limite, une seconde approche a été introduite, la *sélection forward*. On part d'un modèle nul où aucun SNP n'est sélectionné et on rajoute, au fur et à mesure, des SNP au modèle, du plus important au moins important, en fonction de leur contribution à améliorer l'explication du phénotype, avec un critère d'arrêt du type : plus aucun SNP ne contribue significativement à l'amélioration de l'explication de la maladie. Là encore, cette approche est loin d'être optimale. D'une part, elle prend beaucoup de temps quand le nombre de SNP est important (ce qui est le cas dans les études d'association pangénomiques). D'autre part, elle n'aboutit pas forcément à la combinaison optimale, c'est-à-dire qui explique le mieux, notre variable de sortie.

L'approche multivariée ou multi-marqueurs a donc été introduite par les statisticiens dans le but de prendre en compte les relations de dépendances qui existent entre les marqueurs dans l'identification des marqueurs associés au trait phénotypique étudié, et ce parmi un grand nombre de variables (cas de la grande dimension).
Une des approches multivariées les plus performantes, de plus en plus utilisée par les statisticiens est l'approche par pénalisation.

Cette section s'organise en deux grands axes. Dans une première sous-section, nous introduisons l'approche par pénalisation et comparons les différents modèles de régres-

sions avec pénalisation. Ensuite, nous analyserons plus en détails un de ces modèles, celui qui nous a permis de proposer par la suite une approche originale de pénalisation, qu'on expose à la fin.

5.4.1 La méthode pénalisation

Dans le cadre des notations introduites précédemment, rappelons l'expression de l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{ls}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{X_i}\boldsymbol{\beta})^2.$$

On peut citer trois raisons pour lesquelles on demeure non satisfait de cet estimateur des moindres carrés :

- la précision de l'estimation : l'estimateur des moindres carrés a un petit biais mais cependant une variance importante. Ainsi, on peut améliorer la précision de la prédiction en réduisant quelques coefficients et mettant à zéro d'autres pour réduire la variance des valeurs prédites.
- l'interprétation du modèle : on a souvent besoin de faire ressortir un sous-ensemble de variables explicatives pour pouvoir mettre en évidence les effets les plus importants sur la variable à expliquer. En d'autres termes, on cherche un estimateur qui soit **parcimonieux**, c'est-à-dire avec beaucoup de coefficients à 0 et donc un sous-ensemble coefficients β_j non nuls petit par rapport au nombre de marqueurs initial.
- cet estimateur n'est défini que quand le nombre de variables est inférieur au nombre d'observations. En effet, dans le cas contraire, la matrice $\mathbf{^tXX}$ est au plus de rang N et donc pas inversible. Or, le cas $N > p$ est contraire au cadre de la grande dimension de nos études d'association.

L'approche donnée par la méthode de pénalisation consiste à estimer le vecteur de paramètres β par le critère

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{pen}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{X_i}\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||, \tag{5.3}$$

où $||\boldsymbol{\beta}||$ désigne une norme de $\boldsymbol{\beta}$ et λ, un paramètre de régularisation.

5.4.2 Expressions des différents modèles à comparer

Le LASSO

Introduit par Tibshirani [3], l'*estimateur LASSO* (pour *Least Absolute Shrinkage Selection Operator*) s'écrit comme suit :

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{l1}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{X_i}\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||_1, \tag{5.4}$$

$$\text{avec} \qquad ||\boldsymbol{\beta}||_1 = \sum_{j=0}^{p} |\beta_j|.$$

Le LASSO utilise donc une pénalité sur la norme 1 des coefficients de l'estimateur. Le paramètre λ contrôle la parcimonie du modèle, de manière à ce que, si $\lambda \to \infty$, aucun marqueur n'est sélectionné. Et pour une valeur de λ assez petite, tous les marqueurs sont inclus dans le modèle (ont un coefficient non nul)(Voir figure 5.3).
La valeur optimale de λ est celle qui minimise l'erreur de prédiction et peut être calculée à l'aide d'algorithmes tels que la validation croisée.

Pour mieux comprendre le rôle de la pénalisation et le fonctionnement de la méthode LASSO, plaçons nous dans le cas $N > p$ et l'estimateur des moindres carrés est alors bien défini ; et regardons de plus près la forme de l'estimateur dans le cas simple où les marqueurs sont décorrélés, c'est-à-dire dans le cas où la matrice $\mathbf{X}$ est orthogonale.
Dans ce cas, pour $j \in [1, p]$, l'estimateur est de la forme :

$$\hat{\beta}_j^{l1} = sign(\beta_j^{ls})(|\beta_j^{ls}| - \lambda)_+$$

où on a noté β_j^{ls} la solution du modèle en absence de régularisation.

On constate donc que les coefficients β_j^{ls} des moindres carrés se retrouvent **seuillés**. Ils sont rétrécis d'un coefficient λ s'ils sont supérieurs à λ et annulés sinon.
Là encore, cet exemple illustre le rôle du coefficient λ qui détermine, en quelque sorte, la parcimonie de notre modèle : plus λ est grand, plus il y a de coefficients nuls dans notre estimateur.
La pénalisation permet donc une sélection automatique des marqueurs selon un certain critère.

Enfin, l'estimateur LASSO présente une sorte de compromis entre, avoir une bonne vraisemblance et un bon ajustement aux données à travers le premier terme des moindres carrées ; et ne pas sélectionner un grand nombre de variables, en pénalisant la norme 1 des coefficients de régression, à travers le deuxième terme.

Du point de vue algorithmique, la méthode LASSO est implémentée dans le package R : *glmnet*. La fonction du même nom prend comme arguments la matrice $\mathbf{X}$ des génotypes des SNP et un vecteur $\mathbf{y}$ des phénotypes de chaque individu et calcule, pour un certain nombre de valeurs de λ (un *chemin* des λ), les coefficients β relatifs à chaque SNP.
Voici, par exemple, une représentation du résultat de l'algorithme LASSO que nous avons appliqué à des données simulées :

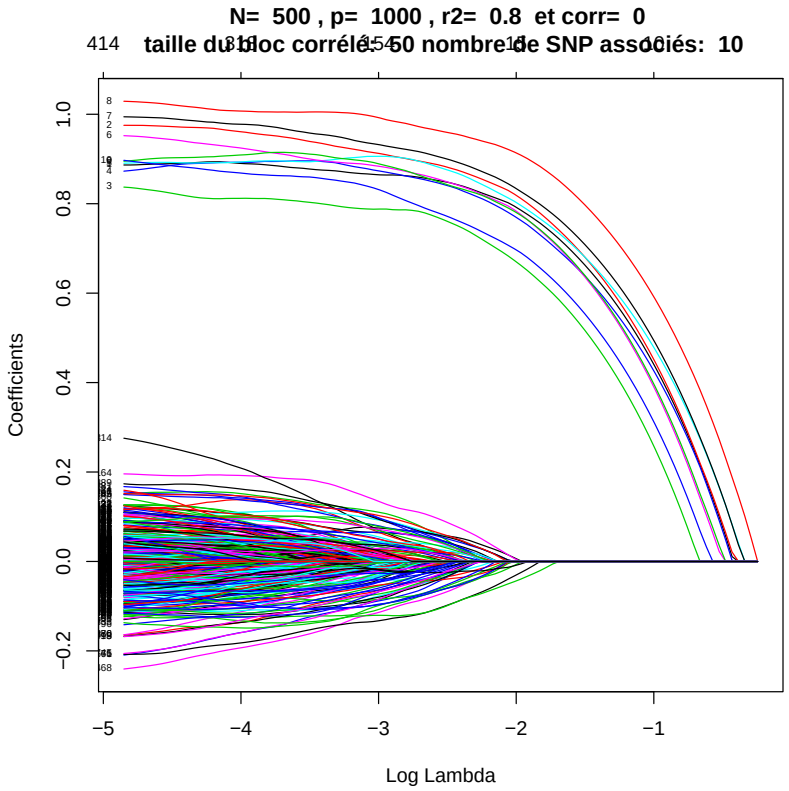


FIGURE 5.3 – Estimation des coefficients en fonction de Log(λ)

La figure 5.3 montre les estimations des coefficients de régression associés à chaque SNP en fonction de Log(λ). On constate que pour une valeur Log(λ)< -1.5, un grand nombre de marqueurs sont sélectionnés par le modèle et donc avec un nombre de faux positifs assez important. Par contre, à partir de la valeur -1.5 de Log(λ), seuls les 10 SNP réellement associés au phénotype sont sélectionnés (avec des coefficients non nuls).
Ainsi, pour un problème assez facile ($r^2 = 0.8$) et pour $N < p$, la méthode LASSO arrive parfaitement à retrouver les vrais positifs à partir d'une certaine valeur seuil de λ. Voici, la figure qu'on obtient, avec les mêmes paramètres de simulation mais pour $r^2 = 0.2$:

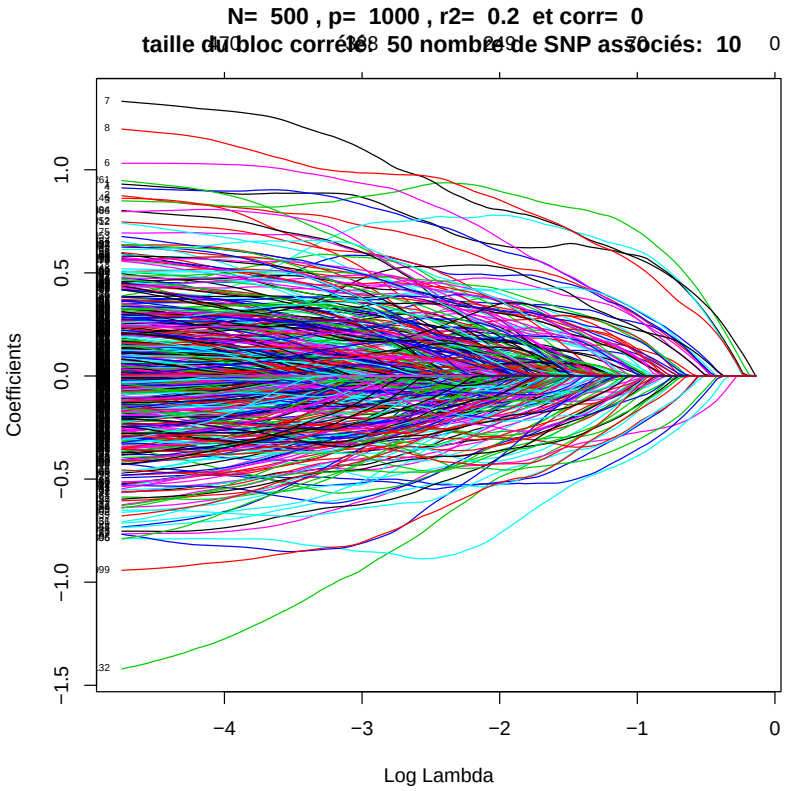
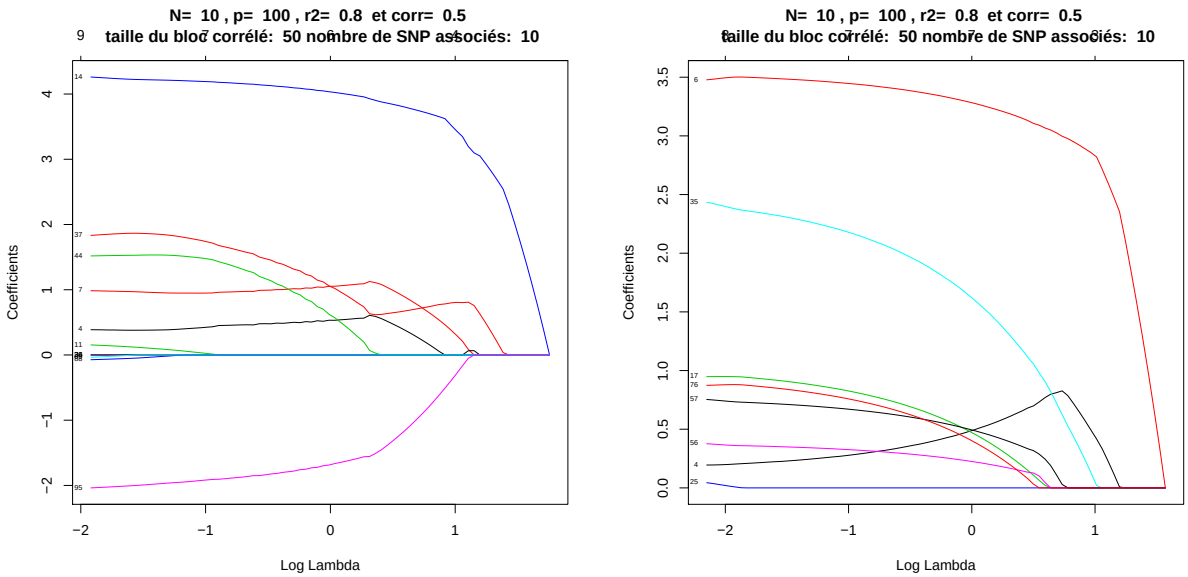


FIGURE 5.4 – Estimation des coefficients en fonction de Log(λ)

On constate que, dans le cas d'un problème difficile, un très grand nombre de coefficients sont non nuls jusqu'à une valeur de λ assez importante. Ainsi, à aucune étape du chemin le LASSO n'arrive vraiment à ressortir un petit nombre de marqueurs significatifs pour expliquer le phénotype.

Nous allons à présent mettre en évidence **une des limites** de cette méthode LASSO. Pour ceci, nous avons effectué deux simulations dans les mêmes conditions : 10 SNP parmi 100 sont associés au phénotype et les *50 premiers* SNP sont corrélés entre eux avec un coefficient de corrélation de 0.5.

Les deux figures 5.5a et 5.5b montrent les résultats obtenus. Pour les deux simulations, à partir d'une certaine valeur de λ, un seul marqueur est sélectionné, celui qui a déjà un coefficient assez important dès le début par rapport aux autres marqueurs. Pour la première simulation, il s'agit du marqueur numéro 14 et pour la deuxième simulation, c'est le marqueur numéro 6. Le marqueur 14 n'est pas directement associé au phénotype mais est corrélé avec des marqueurs qui le sont.
Ainsi, on constate qu'en présence d'un bloc de SNP corrélés (les 50 premiers), le LASSO a tendance à sélectionner un SNP *au hasard*, pas forcément celui qui est réellement associé au phénotype (comme dans le cas du SNP 14), et d'annuler tous les autres SNP du bloc.



(a) Première simulation

(b) Deuxième simulation

L'Elastic-Net

Ce modèle a été proposé par Zou et Hastie [4] pour pallier la limite du modèle LASSO citée précédemment.

Le modèle *Elastic-Net* s'écrit comme suit :

$$\hat{\boldsymbol{\beta}}^{EN} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{X_i}\boldsymbol{\beta})^2 + \alpha||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2, \tag{5.5}$$

$$\text{avec} \qquad ||\boldsymbol{\beta}||_2^2 = \sum_{j=0}^{p} \beta_j^2 \quad \text{et} \quad \alpha \in [0,1].$$

Le rajout de la pénalité en norme 2 permet, dans ce cas, de seuiller tous les coefficients "de la même manière", c'est-à-dire les coefficients de régression auront des valeurs assez proches.

La méthode Elastic-Net est également implémentée dans le même package R que le LASSO (*glmnet*) mais en rajoutant la variable d'entrée α.

CAR Scores

Cette méthode a été proposée par Zuber et Strimmer [5]. Elle consiste à attribuer à chaque marqueur un score de la forme

$$w_j = \mathbf{P}_{j.}^{-1/2}\mathbf{P}_{\mathbf{Xy}} \quad \text{pour tout} \quad j \in [1,p], \tag{5.6}$$

en notant $\mathbf{P}_{j.}$ la $j^{ème}$ ligne de la matrice des corrélations des prédicteurs et $\mathbf{P}_{\mathbf{Xy}}$ le vecteur des corrélations marginales entre la variable à expliquer et les prédicteurs.
Plus le score du marqueur est grand, plus la probabilité que ce dernier soit associé au

phénotype est grande.
Cette approche des CAR Scores est un peu particulière : elle est, en quelque sorte, intermédiaire entre une approche simple marqueur (attribution de scores) et une approche multivariée en exploitant explicitement les relations entre les variables grâce à la matrice **P**.

Nous analyserons, plus en détail, la théorie et les performances de cette méthode plus loin dans le rapport.

5.4.3 Comparaison de méthodes pour la sélection

Le problème de comparaison des méthodes citées précédemment, en termes de sélection, peut être abordé comme un problème de classification binaire. En effet, une variable sélectionnée par une méthode donnée (classée positive) est considérée comme "vrai positif" si elle est réellement associée au phénotype et est un "faux positif" si elle ne l'est pas. En d'autres termes, pour un jeu de données simulé (donc les marqueurs associés au phénotype connus à l'avance), on cherche à comparer la capacité de chaque méthode à sélectionner les "bons" marqueurs, et le plus rapidement possible.

Il existe une représentation graphique très pratique pour comparer des méthodes en classification binaire : les courbes ROC pour *Receiver Operating Characteristic* en anglais. Il s'agit de représenter le taux de vrais positifs (TPR=True Positive Rate) en fonction du taux de faux positifs (FPR=False Positive Rate). Ces deux grandeurs se calculent comme suit :

$$\begin{aligned} TPR &= TP/(TP+FN) \\ FPR &= FP/(FP+TN) \end{aligned}$$

en notant : TP : le nombre de vrais positifs, TN : le nombre de vrais négatifs
FP : le nombre de faux positifs, FN : le nombre de faux négatifs

Par conséquent, pour évaluer les performances d'une méthode donnée en classification, on évalue les TPR et FPR pour différentes valeurs des paramètres de la méthode, ce qui permet de construire une courbe. Ensuite, on peut dire que, plus la surface en dessous de sa courbe ROC est importante, plus la méthode est performante. En effet, la première bissectrice représente la courbe du classificateur binaire aléatoire, qui ne distingue pas réellement les vrais positifs des faux positifs. Ainsi, plus une courbe ROC s'écarte de la première bissectrice, plus elle sélectionne de vrais positifs plus rapidement et donc, plus elle est performante.

Enfin, pour les différentes méthodes citées précédemment et dont on va comparer les performances, nous avons effectué une grille des valeurs du paramètre à faire varier pour chaque méthode mais ces paramètres ne sont pas les mêmes pour toutes les approches. Pour les méthodes LASSO et Elastic-Net, nous avons utilisé une grille de valeurs du paramètre de régularisation (λ pour le LASSO α pour l'Elastic-Net). Concernant les méthodes Univariée et CAR Scores, les valeurs des FPR et TPR ont été calculées pour chaque niveau du test.

5.4.4 Résultats en sélection : en absence de corrélations

En termes de sélection, nous avons représenté les courbes ROC des cinq méthodes : simple-marqueur (ou univariée), CAR Scores, LASSO, Elastic-Net avec α=0.2 et Elastic-Net avec α=0.8. Les paramètres de nos simulations sont les suivants : N (nombre d'individus), p (nombre de SNP), corr (corrélation entre SNP), r2 (difficulté du problème), le nombre de SNP à retrouver et la taille du bloc de SNP corrélés.

Suite à plusieurs simulations en changeant plusieurs paramètres, nous avons constaté qu'en absence de corrélations entre les SNP, l'allure des courbes change principalement en fonction des deux grandeurs N/p et $||\boldsymbol{\beta}||_0/p$ (en posant $||\boldsymbol{\beta}||_0$ le nombre de coefficients de régression non nuls). Nous avons alors effectué les trois simulations suivantes dont les résultats sont présentés dans les figures 5.5, 5.6 et 5.7.

Simulation	N	p	$\|\|\boldsymbol{\beta}\|\|_0$	N/p	$\|\|\boldsymbol{\beta}\|\|_0/p$
Simulation 1	50	100	10	50%	10%
Simulation 2	30	100	10	30%	10%
Simulation 3	100	1000	10	10%	1%

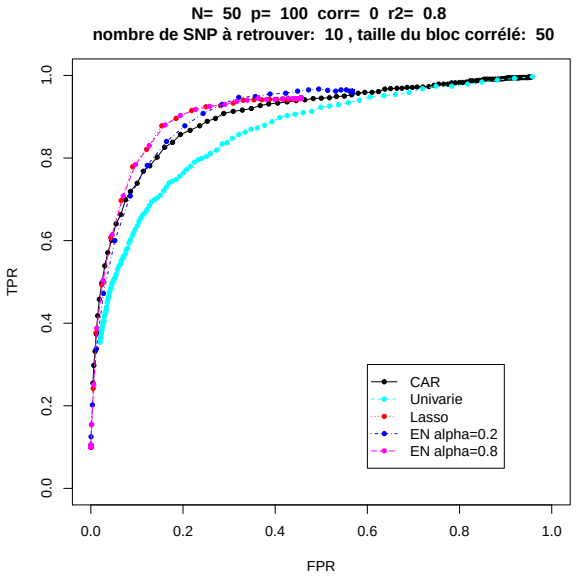


FIGURE 5.5 – simulation 1

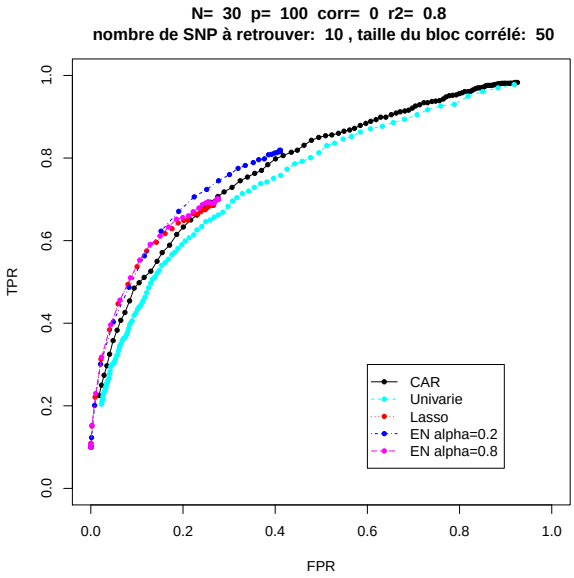


FIGURE 5.6 – simulation 2

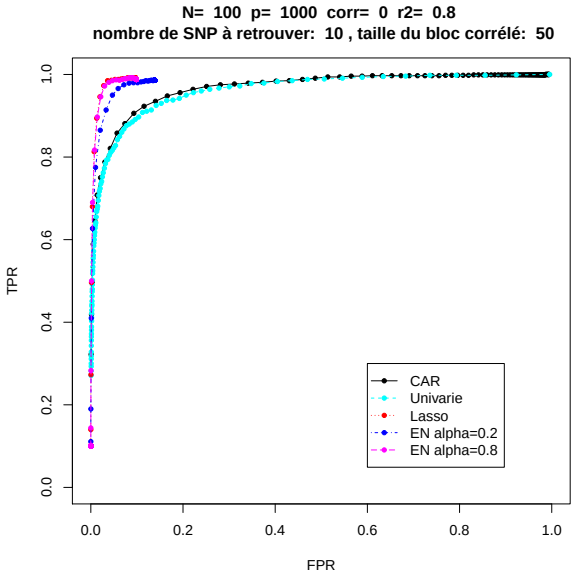


FIGURE 5.7 – simulation 3

D'abord, on constate que, dans les trois cas, les méthodes de régression avec pénalisation sont plus performantes que les méthodes CAR Scores et Univarié. Ceci s'explique par le fait que les méthodes multivariées prennent en compte tous les SNP en un temps et pas un à un individuellement et indépendamment de tous les autres.

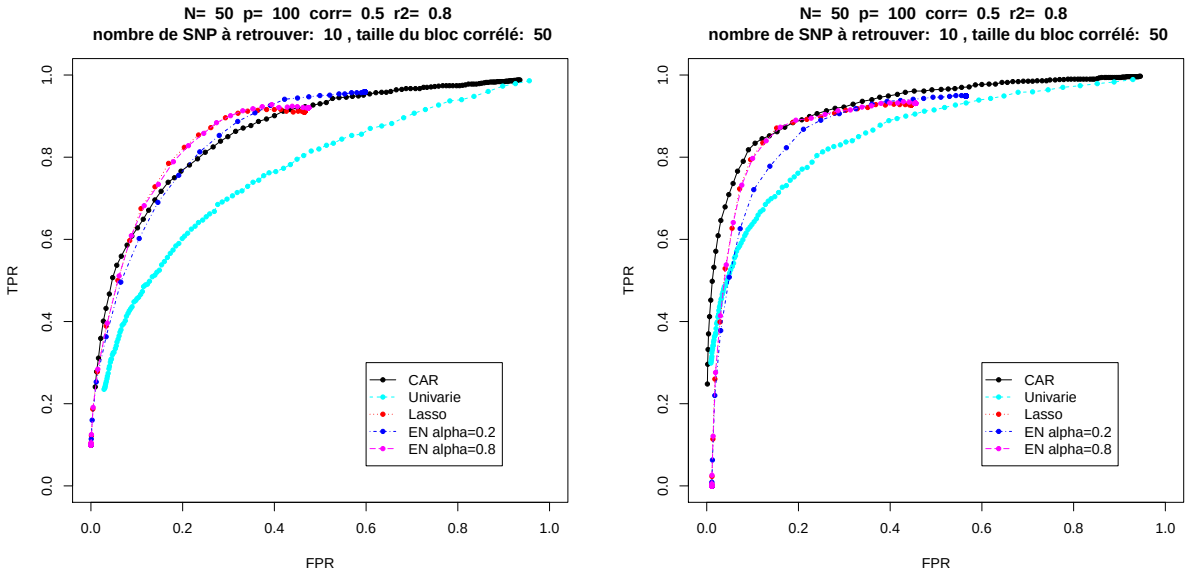
La comparaison des deux premières simulations montre que la diminution du paramètre N/p dégrade la qualité de la sélection pour toutes les méthodes. En effet, le passage d'un rapport N/p de 50% à 30% entraîne un rapprochement de toutes les courbes de la première bissectrice. Cette diminution des performances de toutes les méthodes est due

aux limites mathématiques des modèles de régression en général. Effectivement, pour un nombre d'individus inférieur ou égal au nombre de variables, toutes les méthodes ne peuvent sélectionner plus de N variables. Donc, à p fixé, plus N est petit, moins la méthode est performante vu qu'elle doit retrouver les bons marqueurs tout en sélectionnant un nombre moins important.

La troisième simulation représente le cas où le rapport N/p est faible (10%) et le rapport $||\boldsymbol{\beta}||_0/p$ est très faible (1%). On remarque que les méthodes de pénalisation sont meilleures en termes de sélection puisque leur courbe commence par augmenter verticalement jusqu'à atteindre 1. Ce qui veut dire que ces méthodes commencent par sélectionner tous les vrais positifs avant de rajouter des faux positifs.
Les bonnes performances des méthodes LASSO et Elastic-Net s'expliquent par le fait que la troisième simulation représente, en réalité, le cas de la grande dimension : un nombre de marqueurs très supérieur au nombre d'individus et parmi ce grand nombre de variables, on veut en sélectionner très peu, les plus significatifs. Ainsi, comme on l'a déjà expliqué précédemment, on voit bien que les modèles de régression avec pénalisation sont les plus adaptés au cas de la grande dimension vu qu'il s'agit de modèles de régression multivariée et qui font, en même temps, un travail de sélection de variable grâce au terme de pénalisation.

5.4.5 Résultats en sélection : en présence de corrélations

Dans un deuxième temps, nous avons comparé les mêmes méthodes, pour la sélection, mais en présence de corrélations. Pour des raisons de clarté des figures, nous avons choisi de faire les simulations avec les corrélations avec les paramètres suivants : $N/p = \mathbf{50}\%$, $||\boldsymbol{\beta}||_0/p = \mathbf{10}\%$ et un coefficient de corrélation de 0.5.



(a) Tous les SNP associés au phénotype sont dans le bloc de SNP corrélés

(b) Aucun SNP associé au phénotype n'est dans le bloc des SNP corrélés

Les résultats de la comparaison sont présentés dans les figures 5.8a et 5.8b. Pour la figure de gauche, les simulations ont été faites dans le cas où les 10 SNP associés au phénotype font partie du bloc de 50 SNP corrélés. Pour la figure de droite, les 10 SNP associés au phénotype ne font pas partie du bloc de 50 SNP corrélés.
Dans un premier temps, on peut remarquer que, toutes les méthodes sont, de façon générale, moins performantes dans le cas où les SNP associés au phénotype sont dans le bloc des SNP corrélés. Ce qui est prévisible puisque, dans le cas où les vrais positifs sont corrélés avec de vrais négatifs, toutes les méthodes ont tendance à sélectionner les SNP corrélés avec les vrais SNP associés.
D'autre part, on constate, une fois de plus, que l'approche univariée est la moins performante de toutes, dans les deux cas. Pour ce qui est des méthodes CAR et régression pénalisée, on peut dire qu'elles sont à peu près à performances comparables dans les deux cas. Cependant, dans le cas où les SNP associés au phénotype ne sont pas dans le bloc des SNP corrélés, la méthode CAR semble un peu plus efficace que toutes les autres.

5.5 Analyse de la méthode des CAR Scores et idée de nouvelle méthode

Rappelons la définition du vecteur des CAR Scores :

$$\mathbf{w} = \mathbf{P}^{-1/2}\mathbf{P}_{\mathbf{Xy}},$$

en notant $\mathbf{P}$ la matrice des corrélations des prédicteurs et $\mathbf{P}_{\mathbf{Xy}}$ le vecteur des corrélations marginales entre la variable à expliquer et les prédicteurs.
On peut également l'écrire comme suit :

$$\begin{aligned}\mathbf{w} &= \mathbf{P}^{-1/2}({}^t\mathbf{X}_{std}\mathbf{y}) \\ &=^t (\mathbf{X}_{std}\mathbf{P}^{-1/2})\mathbf{y} \\ &= Corr(\mathbf{X}_{std}\mathbf{P}^{-1/2}, \mathbf{y}),\end{aligned}$$

en notant $\mathbf{X}_{std}$ la matrice $\mathbf{X}$ avec les variables des colonnes centrées et réduites.

Ainsi, les CAR Scores représentent les corrélations entre la réponse phénotypique et les variables explicatives **décorrélées**.
Suite à cette interprétation des CAR Scores, expliquée dans l'article [5], nous avons eu l'idée d'introduire le modèle de régression original suivant, le New LASSO :

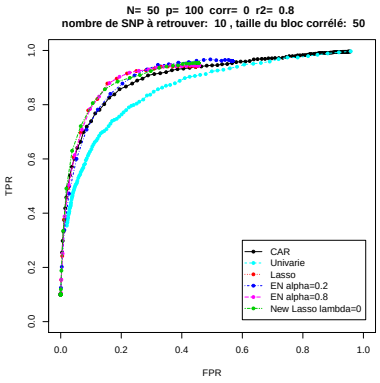
$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\boldsymbol{new}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}((1-\lambda)^{\mathbf{t}}\mathbf{X}\mathbf{X} + \lambda\mathbf{I}_p)^{-1/2}\boldsymbol{\beta}||_2^2 + \gamma||\boldsymbol{\beta}||_1 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}\mathbf{P}_\lambda{}^{-1/2}\boldsymbol{\beta}||_2^2 + \gamma||\boldsymbol{\beta}||_1,\end{aligned} \tag{5.7}$$

où λ et γ sont deux paramètres de régularisation.

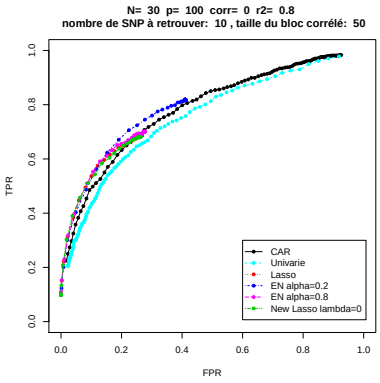
Par rapport aux autres modèles, les avantages mis en évidence par cette nouvelle approche sont les suivants :

- contrairement à la méthode Elastic-Net, avec le New LASSO, la décorrélation se fait de manière directe sur les données, grâce au paramètre λ.
- utiliser l'approche par pénalisation qui, d'après les résultats précédents, est la plus performante **en absence de corrélations**. En effet, cette nouvelle approche est une application du LASSO à de nouvelles entrées qui sont $\mathbf{y}$ et $\mathbf{X}\mathbf{P}_\lambda{}^{-1/2}$ (c'est-à-dire les variables explicatives décorrélées) et dont le paramètre de régularisation est γ.

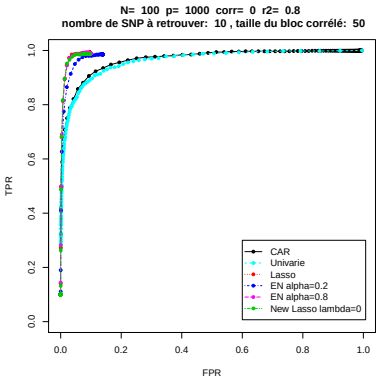
Dans le but de comparer la nouvelle méthode New LASSO avec les autres méthodes existantes, toutes les simulations présentées précédemment ont été réalisées à nouveau, avec les mêmes paramètres, mais en rajoutant la nouvelle méthode.



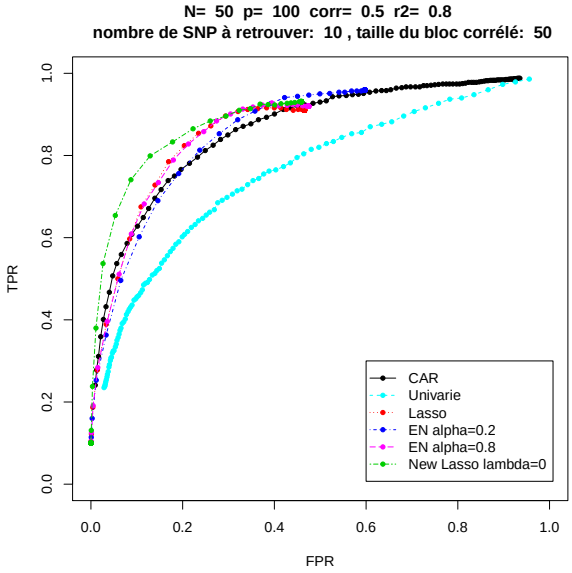
(c) simulation 1



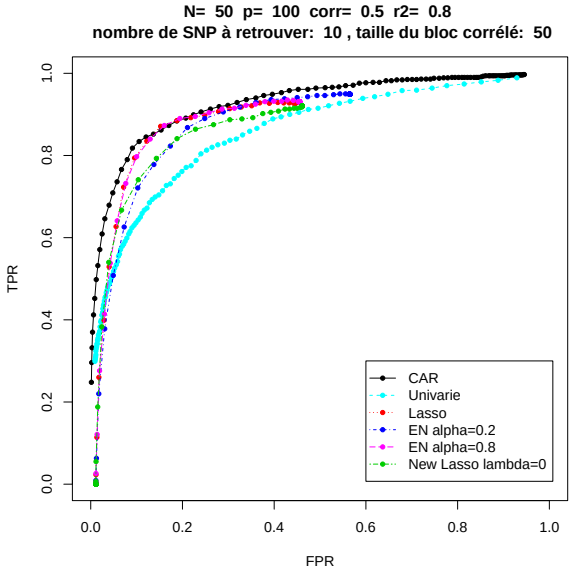
(d) simulation 2



(e) simulation 3



(f) Tous les SNP associés au phénotype sont dans le bloc de SNP corrélés



(g) Aucun SNP associé au phénotype n'est dans le bloc des SNP corrélés

Pour les figures de 5.8c à 5.8g, la nouvelle méthode New LASSO a été rajoutée en vert.

D'abord, plusieurs simulations, en présence et en absence de corrélations, ont permis de constater que la variation du paramètre de régularisation λ n'influe pas de façon considérable sur les performances de la méthode et sa courbe ROC est presque inchangée. Ce paramètre a donc été maintenu à zéro pour toutes les simulations.
Ensuite, concernant les trois simulations en absence de corrélations (figures de 5.8c à 5.8e), on constate que New LASSO est comparable aux méthodes de régression avec pénalisation. Sa courbe ROC est superposée à celle du LASSO dans les trois cas. Ce résultat est prévisible vu qu'en absence de corrélations, le rajout du terme de décorrélation $({}^{\mathbf{t}}\mathbf{XX})^{-1/2}$ devant $\mathbf{X}$ est très proche de l'identité puisque les variables sont déjà décorrélées. Et le modèle New LASSO correspond alors au modèle LASSO dans ce cas.

Le cas le plus intéressant à étudier est effectivement le cas où certains marqueurs sont

corrélés. Comme précédemment, le premier cas correspond au cas où les 10 SNP associés au phénotype font partie du bloc de 50 SNP corrélés. La figure 5.8f montre que la méthode proposée est plus performante que toutes les autres. Comme prévu, les approches avec pénalisation comme LASSO et Elastic-Net ont été dépassées par New LASSO grâce à la décorrélation des données intégrée dans ce dernier. Cette décorrélation améliore considérablement les performances des modèles multivariés avec pénalisation.
Dans la deuxième configuration où les 10 SNP associés au phénotype ne font pas partie du bloc des 50 SNP corrélés, toutes les méthodes existantes s'avèrent un peu plus performantes, comme on l'a constaté précédemment, sauf la méthode New LASSO. La courbe de cette dernière s'approche légèrement de la première bissectrice par rapport à la figure 5.8f. Une interprétation serait que, dans le cas où aucun SNP associé au phénotype n'est dans le bloc des SNP corrélés, la décorrélation effectuée par le New LASSO n'est pas nécessaire. Elle peut même être nuisible en influençant sur les variables réellement associées au phénotype.

Chapitre 6

Bilans

6.1 Organisation et communication

Mon intégration au sein du laboratoire s'est très bien passée. Dès le premier jour, Professeur Christophe Ambroise m'a reçue pour m'expliquer quelques points importants : l'organisation du laboratoire, le déroulement du travail, etc. Il m'a également présentée aux membres de l'équipe avec laquelle je travaillerai : Pierre NEUVIAL, mon second encadrant de stage, Julien CHIQUET et Cyril DALMASSO. Enfin, il m'a fait un résumé des différents objectifs du stage et m'a expliqué les premières tâches à faire pour commencer.

La voie d'approfondissement (VAP) que j'ai suivie à Télécom SudParis a été "Modélisation Statistique et Applications". Et la biostatistique ne faisait pas partie du programme de cette VAP. De ce fait, effectuer mon stage au Laboratoire Statistique et Génome a été pour moi une découverte d'un domaine complètement nouveau des statistiques appliquées. En effet, seulement des généralités sur les modèles mathématiques de régression linéaire ont été vues en début de deuxième année à Télécom SudParis et le langage de programmation utilisé pour les études statistiques en école était Matlab et non le langage R.

Les premières tâches que m'a confié mon encadrant on été les suivantes : pendant les deux premières semaines, alterner entre théorie mathématique sur les modèles de régression multivariée à lire dans le livre [6] et la pratique en lisant un tutoriel sur le langage R et faire de petits exercices pratiques disponibles sur le site du laboratoire. Ensuite, pendant tout le stage, mes encadrants étaient toujours disponibles sur place ou par e-mail, pour répondre à toutes mes questions. Des réunions avec toute l'équipe étaient organisées très régulièrement pour rendre compte de l'avancement des travaux ou des réflexions sur l'orientation des recherches.
Par moments, les résultats des simulations n'étaient pas très concluants ; et des réunions plus fréquentes étaient alors organisées pour se pencher, dans un premier temps sur le code ou ensuite essayer de nouvelles pistes de recherche si les résultats ne concordaient pas vraiment avec la théorie mathématique.

En ce qui concerne les possibilités de thèse au laboratoire, la stratégie de ce dernier est de présenter, chaque année, un unique candidat au concours pour les contrats doctoraux

organisé par l'Ecole Doctorale "Des Génomes Aux Organismes" (GAO), dans le but de maximiser les chances d'obtention de bourse doctorale. Cette année, mon dossier a été présenté et des auditions pour les présentations des projets de thèses ont été organisées les 3 et 4 Juillet. Ainsi, tout le mois de Juin a été consacré à la préparation de l'audition.

Mes encadrants m'ont d'abord beaucoup aidée à définir et exposer clairement le sujet de thèse. Sachant que le jury, constitué d'une quinzaine de personnes, était presque constitué exclusivement de biologistes, il fallait donc, pour ma présentation, mettre l'accent sur l'intérêt biologique de mon projet de thèse et simplifier le plus possible l'aspect mathématique du sujet. De plus, pendant toute la dernière semaine du mois de Juin, j'ai répété tous les jours ma présentation devant mes encadrants et tous les membres du laboratoire qui voulaient bien me donner quelques derniers conseils. Après chaque répétition, je refaisais de petites modifications en prenant en compte les remarques des uns et des autres afin que cela soit le plus clair possible devant le jury du 3 Juillet. Les remarques et le soutien de tous m'ont été d'une grande aide.

6.2 Bilan général

J'ai choisi de faire mon stage de fin d'études dans le domaine des statistiques appliquées à la biologie au Laboratoire Statistique et Génome. Dans ce projet de fin d'études, la problématique est la recherche de marqueurs associés à un phénotype en grande dimension. Une première partie du stage a été consacrée à la compréhension, l'implémentation et la comparaison des méthodes existantes en sélection de variables comme les modèles Univarié, LASSO, Elastic-Net et CAR Scores ; mettre en évidence les avantages et les inconvénients de chacun, pour pouvoir, dans un deuxième temps, proposer la nouvelle méthode New Lasso qui a également ses avantages et ses limites.

Concernant les objectifs scientifiques fixés au début du stage, cette première partie constituera une grande partie bibliographique de mon futur travail de thèse.

Le modèle de régression original proposé, le New LASSO, ne répond pas à la problématique posée au début du stage, à savoir la prise en compte de la forte structuration des données génomiques due aux phénomènes biologiques comme le déséquilibre gamétique ou les interactions entre SNP. En effet, le New LASSO pourrait peut être amélioré en rajoutant l'information sur la structure des données. Le travail sur ce point commencera donc pendant la thèse.

Au regard des relations qui existent entre les marqueurs et les loci de susceptibilité, à savoir les associations résultant principalement du phénomène de déséquilibre gamétique, les approches simple-marqueur apparaissent limitées pour prendre en compte la complexité des structures des données. Les statisticiens se sont donc tournés vers les approches multi-marqueurs qui présentent, eux aussi, des limites méthodologiques liés à la grande dimension des données.
Les résultats de ce projet de fin d'études étant présentés en deux parties -l'approche univariée et l'approche multivariée-, ces deux approches peuvent paraître au premier abord

indépendantes. En réalité, la stratégie multi-marqueurs dépend énormément de l'approche simple-marqueur. En effet, une grande partie des méthodes multi-marqueurs ne peuvent pas gérer le très grand nombre de données à traiter. Ainsi, une analyse simple-marqueur peut être effectuée comme première étape de pré-sélection de marqueurs. De ce fait, la pertinence et la validité de la méthode univariée influe considérablement sur l'efficacité des approches multivariées.

6.3 Bilan personnel

Du point de vue scientifique, les compétences et les connaissances que j'ai acquises sont incontestables : des lectures d'articles scientifiques sur les modèles de régression les plus récents à l'apprentissage et le développement de programmes en langage R en passant par la découverte du monde de la recherche à travers les séminaires. D'une part, j'ai approfondi les connaissances en statistiques que j'ai acquises en suivant la voie d'approfondissement "Modélisation Statistique et Applications" à Télécom SudParis. J'ai, d'autre part, découvert un nouveau domaine d'application des statistiques qui est la génomique à travers la découverte et la maîtrise des modèles de régression utilisés dans les études d'association pangénomiques. Il s'agit d'un domaine pluridisciplinaire et très vaste et qui nécessite des compétences en biologie, en informatique et en mathématiques. C'est ce qui explique la période assez longue de trois mois que j'ai pris, au début de mon stage, pour comprendre les notions de base en biologie qui concernent mon sujet de stage, les modèles mathématiques qui y sont liés et le langage de programmation pour les simulations.

Mon séjour au Laboratoire Statistique et Génome m'a permis d'approcher le monde de la recherche scientifique et travailler pendant six mois au sein d'une vraie équipe de recherche m'a appris plusieurs choses et m'a permis de développer plusieurs qualités indispensables si je veux continuer dans la recherche.

Grâce à mes encadrants qui m'ont toujours guidée dans l'apprentissage d'une vraie démarche scientifique de recherche de résultats scientifiques, j'ai pu avoir, au début du mois de Mai, un premier programme de comparaison des méthodes de régression existantes. Ayant des idées sur les avantages et les limites de ces approches existantes et plusieurs pistes de méthode originale, mes encadrants m'ont proposé de participer aux Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM) organisées à Rennes du 2 au 7 Juillet, en présentant un poster. J'avais donc trois semaines pour implémenter la nouvelle méthode New LASSO, en espérant avoir des résultats satisfaisants permettant de conclure quant aux performances de chacune et envoyer à l'équipe organisatrice un résumé de ce que je vais exposer dans le poster.
Finalement, je n'ai malheureusement pas pu terminer ce travail en trois semaines. Plusieurs problèmes liés aux simulations ont été soulevés au fur et à mesure que l'on avançait dans l'analyse et l'implémentation de chaque méthode. Cette expérience m'a tout de même appris beaucoup sur le travail en équipe et sur le travail de recherche en général :

- l'importance de la bonne communication au sein de l'équipe et particulièrement pendant les réunions. Il est primordial que tout le monde comprenne la probléma-

tique posée pour pouvoir suggérer des solutions.
- faire en sorte que tous les membres de l'équipe soient au courant de l'avancement du projet en organisant des réunions régulières mais aussi à travers les mails ou les dépôts *svn* pour les codes en R.
- le travail de recherche demande beaucoup de patience et surtout de persévérance. Un bon résultat scientifique est rarement rapide à obtenir et est toujours précédé de plusieurs "fausses pistes" et de résultats préliminaires. Il ne faut pas baisser les bras quand le résultat ne correspond pas à l'intuition première ou que la performance d'une méthode n'est pas si bonne que cela. Il faut commencer par revoir la démarche suivie puis se convaincre que l'intuition n'est pas toujours juste, si la démarche suivie ne présente aucun défaut.

J'ai tout de même participé au séminaire JOBIM où j'ai côtoyé et échangé avec des chercheurs de haut niveau en biostatistique.

Enfin, l'objectif, le plus important peut-être de ces six mois de stage, a été atteint : j'ai pu obtenir une bourse de thèse pour trois ans de l'Ecole Doctorale GAO. J'ai été classée première ex-æquo avec trois autres candidats, parmi les quinze qui se présentaient au concours pour l'attribution des contrats doctoraux.

6.4 Responsabilité sociétale des organismes

La responsabilité sociétale des Organismes publics (RSO) est la contribution de ces derniers aux enjeux de développement durable. Cette responsabilité se traduit donc par un ensemble de mesures que prennent les organismes publics pour la protection de l'environnement et pour réduire l'impact écologique de leurs activités sur ce dernier.

Socialement parlant, ce laboratoire de recherche est très cosmopolite, avec des chercheurs, des doctorants et des stagiaires des quatre coins du monde avec une absence totale d'une quelconque discrimination. Bien au contraire, l'aspect multinational au laboratoire est une caractéristique que ce laboratoire de recherche cherche absolument à garder, conscient de son importance pour l'avancement de ses travaux de recherche et de son rayonnement international.

Sur le plan environnemental, vu qu'il s'agit d'un laboratoire de recherche en mathématiques appliquées, il n'existe pas vraiment de système de management de l'environnement à proprement parler. Le fait que chaque employé n'a besoin que de son ordinateur personnel, de papier et d'un stylo pour travailler, montre que l'impact de l'activité principale de ce laboratoire sur l'environnement est minime et très négligeable devant celui d'autres entreprises.
Enfin, il n'y a pas de direction de développement durable dédiée aux stratégies futures dans ce domaine.

Conclusions

Mon stage au sein du Laboratoire Statistique et Génome a été une expérience enrichissante aussi bien pour moi que pour le laboratoire qui m'a accueillie.

En ce qui me concerne, il m'a permis d'acquérir un grand nombre de compétences techniques, personnelles et professionnelles qui me seront indispensables plus tard, dans ma carrière professionnelle.

En approchant le milieu de la recherche scientifique, cela m'a appris, entre autres choses, le travail en groupe, la rigueur qu'exige tout travail scientifique, la communication, la gestion du temps ainsi que l'adaptation à un nouveau milieu, le milieu des laboratoires scientifiques qui est très différent de celui des entreprises.

Mes compétences de futur ingénieur ont également beaucoup apporté au laboratoire aussi bien à travers les tâches que j'ai effectuées qu'à travers le contrat doctoral que j'ai pu obtenir grâce au projet de thèse que j'ai su m'approprier et présenter de façon convaincante.

Enfin, mon stage au sein du Laboratoire Statistique et Génome m'a permis de côtoyer une équipe compétente et motivée, et surtout de nouer des relations amicales avec toute l'équipe.
L'opportunité de continuer mes travaux en réalisant une thèse au sein du même laboratoire me permettra, d'une part, de continuer à évoluer au sein de cette même équipe avec laquelle l'entente scientifique et personnelle est confirmée et d'autre part, me conforter dans mon choix de faire de la recherche dans le domaine des mathématiques appliquées.

Dès mon entrée à Télécom SudParis, mes ambitions professionnelles étaient orientées vers la recherche scientifique et plus spécifiquement dans le domaine des mathématiques. Le statut d'enseignant-chercheur me permettrait de diversifier mes activités et mes compétences professionnelles : d'une part, apprendre la pédagogie en enseignant les mathématiques à niveau avancé et, d'autre part, continuer les travaux de recherche en mathématiques appliquées à des domaines comme la médecine et la biologie. C'est justement cet aspect pluridisciplinaire qui a orienté mes choix professionnels vers la biostatistique.

Bibliographie

[1] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6) :714–721, 2009.

[2] C. Dalmasso, W. Carpentier, L. Meyer, C. Rouzioux, C. Goujard, M.L. Chaix, O. Lambotte, V. Avettand-Fenoel, S. Le Clerc, L.D. de Senneville, et al. Distinct genetic loci control plasma hiv-rna and cellular hiv-dna levels in hiv-1 infection : the anrs genome wide association 01 study. *PLoS One*, 3(12) :e3907, 2008.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[4] H. Zou and T. Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. *Journal of the Royal Statistical Society : Series B. v67*, pages 301–320, 2003.

[5] V. Zuber and K. Strimmer. High-dimensional regression and variable selection using car scores. *Arxiv preprint arXiv :1007.5516*, 2010.

[6] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

[7] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R.J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 2010.

[8] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7) :499–511, 2010.

[9] M. Slawski, W. Zu Castell, and G. Tutz. Feature selection guided by structural information. *The Annals of Applied Statistics*, 4(2) :1056–1080, 2010.

[10] J. Liu, K. Wang, S. Ma, and J. Huang. Accounting for linkage disequilibrium in genome-wide association studies : A penalized regression method. Technical report, Technical report, Department of Statistics and Actuarial Science, University of Iowa, 2011.

[11] K.L. Ayers and H.J. Cordell. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8) :879–891, 2010.

[12] S Younkin, J Nadeau, R Elston, and J. S. Rao. The linkage disequilibrium LASSO for SNP selection in a genetic association study of late onset alzheimer disease. Technical report, 2010.

[13] S Kim and E P Xing. Exploiting genome structure in association analysis. *J Comput Biol*, May 2011.

[14] A.S. Foulkes. *Applied statistical genetics with R : for population-based association studies*. Springer Verlag, 2009.

[15] V. Zuber, A. Silva, and K. Strimmer. An efficient approach to simultaneous snp selection : A case study on gaw17 data. *Arxiv preprint arXiv :1203.3082*, 2012.

[16] M. Szafranski. Pénalités hiérarchiques pour l'ntégration de connaissances dans les modèles statistiques. 2008.

[17] M. Guedj. Méthodes statistiques pour l'analyse de données génétiques d'association à grande echelle. 2007.