université
PARIS-SACLAY

Lien entre héritabilité et prédiction de phénotypes complexes chez l'humain : une approche du problème par la régression ridge sur des données de population

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 577, Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Sciences de la vie et de la santé
Unité de recherche: CNRGH - Centre National de Recherche en Génomique Humaine
Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Evry, le 24/11/2020, par

Arthur FROUIN

Composition du jury:

Emmanuelle GÉNIN Directrice de recherche INSERM, Université de Bretagne occidentale U1078	Présidente
Anne-Laure BOULESTEIX Professeure de biométrie, Université Ludwig Maximilian de Munich	Rapportrice
David-Alexandre TRÉGOUËT Directeur de la Recherche (DR1), Centre Bordeaux Population Research INSERM U1219	Rapporteur
Benoit LIQUET Professeur des universités, Université de Pau et Pays de l'Adour, UMR CNRS 5142	Examinateur
Hervé PERDRY Maître de conférences, HDR, Université Paris-Sud U1018	Examinateur
Jean-François DELEUZE Directeur du Centre National de Recherche en génomique Humaine, CEA, Evry	Directeur
Christophe AMBROISE Professeur des universités, Université d'Évry Val d'Essonne	Coencadrant
Edith LE FLOCH Ingénieur Chercheur au Centre National de Recherche en génomique Humaine, CEA, Evry	Coencadrante

Thèse de doctorat

NNT: 2020UPASA000

Remerciements

Résumé

Cette thèse étudie l'apport des méthodes d'apprentissage automatique pour la prédiction de phénotypes humains complexes et héritables, à partir de données génétiques en population. En effet, les études d'association à l'échelle du génome (GWAS), qui constituent l'approche la plus classique, n'expliquent en général qu'une petite fraction de l'héritabilité observée sur des données familiales. Cependant l'héritabilité peut être approchée sur des données de population par l'héritabilité génomique, qui estime la variance phénotypique expliquée par l'ensemble des polymorphismes nucléotidiques (SNP) du génome à l'aide de modèles mixtes. Cela suggère qu'un phénotype complexe peut être en partie expliqué par l'addition des petites contributions de nombreux SNP.

Cette thèse aborde donc l'héritabilité du point de vue de l'apprentissage automatique et examine le lien étroit entre les modèles mixtes et la régression ridge.

Notre contribution est double. Premièrement, nous proposons d'estimer l'héritabilité génomique en utilisant une approche prédictive via la régression ridge et la validation croisée généralisée (GCV). Nous montrons que cela est cohérent avec l'estimation basée sur un modèle mixte classique dans le cas de phénotypes quantitatifs.

Deuxièmement, nous dérivons des formules simples qui expriment la précision de la prédiction par la régression ridge en fonction du rapport $\frac{n}{p}$, où n est la taille de la population et p le nombre total de SNP. Ces formules montrent clairement qu'une héritabilité élevée n'implique pas une prédiction précise lorsque $p > n$.

L'estimation de l'héritabilité via GCV et les formules de précision de prédiction sont validées à l'aide de données simulées et de données réelles de UK Biobank.

La dernière partie de la thèse présente des résultats sur des phénotypes qualitatifs. Ces résultats permettent une meilleure compréhension des biais des méthodes d'estimation d'héritabilité.

Abstract

This thesis studies the contribution of machine learning methods for the prediction of complex and heritable human phenotypes, from population genetic data. Indeed, genome-wide association studies (GWAS), which constitute the more classic approach, generally only explain a small fraction of the heritability observed in family data. However, heritability can be approximated on population data by genomic heritability, which estimates the phenotypic variance explained by the set of single nucleotide polymorphisms (SNPs) of the genome using mixed models. This suggests that a complex phenotype can be partly explained by the addition of the small contributions of many SNPs.

This thesis therefore approaches heritability from a machine learning perspective and examines the close link between mixed models and ridge regression.

Our contribution is twofold. First, we propose to estimate genomic heritability using a predictive approach via ridge regression and generalized cross validation (GCV). We show that this is consistent with the estimation based on a classical mixed model in the case of quantitative phenotypes.

Second, we derive simple formulas that express the precision of the ridge regression prediction as a function of the ratio $\frac{n}{p}$, where n is the size of the population and p is the total number of SNPs. These formulas clearly show that a high heritability does not imply an accurate prediction when $p > n$.

Heritability estimation via GCV and prediction precision formulas are validated using simulated data and real data from UK Biobank.

The last part of the thesis presents results on qualitative phenotypes. These results allow a better understanding of the biases of the heritability estimation methods.

Valorisation scientifique

Communications Orales

— European Mathematic Genetic Meeting (2018). Quantifying heritability through a prediction measure.

— Journées de la Statistique (2018). Quantifier l'héritabilité génomique via une mesure de prédiction.

— European Mathematic Genetic Meeting (2019). Quantifying heritability with ridge regression.

Publication

— Frouin A, Dandine-Roulland C, Pierre-Jean M, Deleuze J-F, Ambroise C, Le Floch E. High heritability does not imply accurate prediction under the small additive effects hypothesis. arXiv :200705424 [q-bio, stat] [Internet]. 13 juill 2020 ; Disponible sur : http ://arxiv.org/abs/2007.05424 ; Soumis dans Frontiers in Genetics.

Table des matières

Résumé **ii**

Abstract **iii**

Communication Orales iv

Publications iv

Valorisation scientifique **iv**

Table des matières **v**

Liste des figures **x**

Liste des tableaux **xiii**

Introduction **1**

1 Théorie de l'apprentissage **7**

1.1 Convention de notations 7

1.2 Notion d'apprentissage 8

1.2.1 Terminologie de l'apprentissage 8

1.2.2 Perte d'erreur quadratique et fonction de régression 8

1.3 Un premier exemple de fonction de régression avec le modèle linéaire . 9

1.4 Concept de sur-apprentissage et un exemple avec l'estimateur des moindres carrés 11

1.5 Concept d'erreur 12

1.6 Sélection et évaluation de modèle 13

2 La régression ridge **17**

2.1 Définition et origine 17

2.2 Standardisation des données 18

2.3 Calcul de l'estimateur 19

2.4 Propriétés de l'estimateur 19
2.4.1 Décomposition en valeurs singulières, analyse en composantes principales et degrés de liberté 19
2.4.2 Le dilemme biais-variance pour la régression ridge 21
2.4.3 Théorème de l'existence 22
2.4.4 Un exemple pratique 22
2.4.5 La régression ridge en grande dimension 22
2.5 Choix du paramètre de pénalisation λ 24
2.5.1 Comment choisir λ ? 24
2.5.2 La validation croisée 26
2.6 Les modèles mixtes 29
2.6.1 Lien entre modèle à effets aléatoires et régression ridge 33
2.7 Intégration de variables non-pénalisées dans la régression ridge 35
2.8 Extension du lien entre régression ridge et modèle à effets aléatoires en présence d'effets fixes 38

3 Validation croisée généralisée en grande dimension 40
3.1 Illustration des problèmes de la GCV en grande dimension et sous centrage empirique 41
3.2 Démonstrations des problèmes 43
3.3 Deux corrections pour la GCV 47
3.3.1 Une correction avec une matrice de contraste 47
3.3.2 Une correction brisant les dépendances avec un deuxième ensemble de données 49

4 Application de la régression ridge à l'estimation d'héritabilité 52
4.1 Notations et définitions des données 52
4.2 L'estimation d'héritabilité par les modèles mixtes 53
4.3 Régression ridge et héritabilités 55
4.3.1 Lien entre h^2 et λ 56
4.3.2 Héritabilité prédictive 57
4.3.3 L'héritabilité comme un ratio de variances 60
4.4 Simulations et comparaisons avec l'approche REML 60
4.4.1 Descriptif des simulations 61
4.4.2 Analyse de l'estimation d'h^2_g 64
4.4.3 Analyse de l'estimation de h^2_p et h^2_r 66

4.5 Application aux données UKBiobank 72
4.5.1 Description des données 73
4.5.2 Description et influence des contrôles qualité 74
4.5.3 Calcul des covariables de structure 76
4.5.4 Prise en compte des covariables 79
4.5.5 Estimations d'h^2 79
4.6 En résumé 81

5 Pouvoir prédictif de la régression ridge 84
5.1 Contexte 84
5.2 Une approximation de pouvoir prédictif selon le rapport $\frac{n}{p}$ pour des données de GWAS 86
5.2.1 Idée de l'approximation 86
5.2.2 Pour l'erreur de prédiction sur l'ensemble de test 87
5.2.3 Pour l'erreur de prédiction sur l'ensemble d'apprentissage 92
5.2.4 Pour le carré de la corrélation 94
5.3 Interprétation de l'approximation 96
5.4 Simulations 97
5.4.1 Description des simulations 97
5.4.2 Résultat pour l'erreur de prédiction sur l'ensemble de test . . . 99
5.4.3 Résultats pour le carré de la corrélation 101
5.5 Application aux données UK Biobank 104
5.5.1 Description de l'approche 104
5.5.2 Évolution des pouvoirs prédictifs selon n sur UKBiobank. 107
5.5.3 Ajustement de notre approximation aux données 116
5.6 En résumé 120

6 L'estimation d'héritabilité pour les phénotypes qualitatifs 121
6.1 Contexte 122
6.1.1 Modèle de liability et calculs de Falconer 122
6.1.2 Calculs d'héritabilité sur des individus non-apparentés 122
6.2 Biais observés pour GCTA 127
6.2.1 Description des simulations 127
6.2.2 Les biais de GCTA dans le cas binaire 128
6.3 Notre contribution 130
6.3.1 Discussion des résultats des auteurs de PCGC 130

6.3.2 Une application aux données de cardiomyopathie 134
6.3.3 Une explication pour les valeurs aberrantes de PCGC 141
6.4 En résumé 144

7 Perspectives **145**
7.1 Un résumé de la thèse en quelques lignes 145
7.2 Distribution de paramètre de pénalisation optimaux pour la régression ridge 146
7.3 Utilisation de la régression ridge hétéroscédastique 147
7.4 Vérification de l'approximation de pouvoir prédictif pour des données réelles. 148
7.5 Quitter le modèle linéaire avec la kernel ridge 149
7.6 Une régression ridge à pénalisation négative 149

Bibliographie **162**

A Preuves pour la régression ridge **I**
A.1 Espérance d'une forme quadratique I
A.2 Décomposition biais-variance de l'erreur de prédiction I
A.3 Théorème de l'existence II
A.4 Élements de preuves pour l'estimateur de la LOO III
A.5 Élements de preuves pour la GCV IV

B Graphes supplémentaires pour l'influence du pourcentage de variant causaux pour l'estimation d'héritabilité avec le R^2 et h_r^2 **VI**

C Preuves et graphes supplémentaires pour l'approximation du pouvoir prédictif **XI**
C.1 Calcul du MSE sur un individu de test XI
C.1.1 Deux égalités XI
C.1.2 Limites des quantités XII
C.2 Calcul du MSE sur l'ensemble d'apprentissage XII
C.3 Calcul du carré de la corrélation sur un individu de test XIV
C.4 Calcul de l'estimateur de la variance des coefficients de $\hat{u}_R$ XVI
C.5 Pourcentage d'individus de l'ensemble d'apprentissage passant les filtres contrôle qualité XVII
C.6 Estimation de pouvoir prédictifs avec paramètre de pénalisation fixé . . XVIII

D Estimation d'héritabilité pour le cas qualitatif **XXIV**

D.1 Démonstration de PCGC XXIV

D.1.1 Calcul de probabilités pour la modélisation cas-contrôle XXIV

D.1.2 L'idée de PCGC XXV

D.1.3 Approximation de $\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}^*_{ij})$ XXV

D.2 Estimation des variances des composantes du modèle XXIX

Liste des figures

1.1 Exemple de surapprentissage avec les moindres carrés 12

2.1 Un exemple de la décomposition biais variance de l'estimateur de la régression ridge. 23
2.2 Un exemple du comportement de $\mathrm{err}_{\mathcal{A}}$ et $\mathrm{Err}_{\mathcal{T}}$. 25
2.3 Principe de la validation croisée. 26

3.1 Un exemple du biais de la GCV. 42
3.2 Un exemple de GCV contrastée. 51

4.1 Découpage de la matrice de données de l'ensemble d'apprentissage . . . 62
4.2 Graphes d'estimation d'héritabilité génomique sur des simulations. . . . 65
4.3 Graphe d'estimation des h^2_r sur ensemble d'apprentissage et de test pour les simulations. 67
4.4 Graphe d'estimation des h^2_p sur ensemble d'apprentissage et de test pour les simulations. 68
4.5 Structure de population dans UK Biobank. 75
4.6 Nombre de variants inclus dans les variables de structure pour chaque population. 78
4.7 Régions à fort DL 78
4.8 Estimation d'héritabilité pour la taille sur UKBiobank. 82
4.9 Estimation d'héritabilité pour l'IMC sur UKBiobank. 82
4.10 Estimation d'héritabilité pour la circonférence des hanches sur UKBiobank. 83
4.11 Estimation d'héritabilité pour le tour de taille sur UKBiobank. 83

5.1 Erreur quadratique, carré du biais et variance théoriques sur l'ensemble de test selon le logarithme du ratio n/p. 93
5.2 Erreur quadratique théorique sur l'ensemble de d'apprentissage selon le logarithme du ratio n/p. 94

5.3 Carré de la corrélation théorique sur l'ensemble de test selon le logarithme du ratio n/p. 95
5.4 Comportement de l'erreur quadratique, du biais et de la variance des simulations selon le ratio n/p 100
5.5 Variabilité de l'erreur quadratique selon le log du ratio n/p 101
5.6 Carré de la corrélation moyen sur un ensemble de test selon le ratio n/p. 103
5.7 Graphes d'estimation d'héritabilité, de d.d.l.e et de variance empirique de $\hat{u}_R$ sur des sous-échantillons de UKBB. 111
5.8 Graphes des MSE estimés sur des sous-échantillons de UKBB avec héritabilité estimée. 112
5.9 Graphes des carrés de la corrélation estimés sur des sous-échantillons de UKBB avec héritabilité estimée. 113
5.10 Graphes des R^2 estimés sur des sous-échantillons de UKBB avec héritabilité estimée. 114
5.11 Graphes des h_r^2 estimées sur des sous-échantillons de UKBB avec héritabilité estimée. 115
5.12 Ajustement de l'approximation sur les données de UKBiobank 117
5.13 Régression pour l'ajustement de l'approximation de l'erreur quadratique aux données de UKBiobank. 119

6.1 Graphe du modèle de liability 123
6.2 Graphe sur l'influence du déséquilibre K et K_e sur la liability. 125
6.3 Biais des méthodes basées sur le REML pour des valeurs de K faibles. . 129
6.4 Biais des méthodes basées sur le REML quand n augmente. 130
6.5 Graphe d'influence de n, h_{sim}^2 et K_e sur l'estimation d'héritabilité pour le cas binaire. 132
6.6 Graphe d'influence de n, h_{sim}^2 et K_e sur la corrélation GxE. 133
6.7 Graphe des deux premières composantes principales pour les données de cardiomyopathie dilatée pour le premier jeu de filtres. 136
6.8 Graphe des deux premières composantes principales pour les données de cardiomyopathie dilatée pour les deux jeux de filtres cumulés. 138
6.9 Graphe sur l'influence des composantes principales pour l'estimation d'h^2 en binaire. 140
6.10 Ensemble de graphes pour illustrer les valeurs aberrantes de PCGC . . 143

B.1 Graphe d'estimation de h_r^2 sur l'ensemble d'apprentissage pour des simulations. VII
B.2 Graphe d'estimation de h_r^2 sur l'ensemble de test sur des simulations. . VIII
B.3 Graphe d'estimation de h_p^2 sur l'ensemble d'apprentissage pour des simulations. IX
B.4 Graphe d'estimation de h_p^2 sur l'ensemble de test sur des simulations. . X

C.1 Portion d'individus passant les filtres XVII
C.2 Graphes d'estimation d'héritabilité, de d.d.l.e et de variance empirique de $\hat{u}_R$ sur des sous-échantillons de UKBB. XIX
C.3 Graphes des MSE estimés sur des sous-échantillons de UKBB avec héritabilité fixée. XX
C.4 Graphes des R^2 estimés sur des sous-échantillons de UKBB avec héritabilité fixée. XXI
C.5 Graphes des h_r^2 estimés sur des sous-échantillons de UKBB avec héritabilité fixée. XXII
C.6 Graphes des corrélations estimés sur des sous-échantillons de UKBB avec héritabilité fixée. XXIII

D.1 Graphe d'influence de n, h_{sim}^2 et K_e sur l'estimation des composantes de variance. XXIX

Liste des tableaux

4.1 Table des valeurs des paramètres pour les simulations 61
4.2 Effet des contrôles qualité sur les différentes populations et approches. . 77

5.1 Taille des ensembles d'apprentissage pour l'évaluation du pouvoir prédictif sur les données réelles. 107
5.2 Nombre de répétitions pour l'évaluation du pouvoir prédictif sur données réelles. 107
5.3 Valeur d'héritabilité fixée pour le comportement des quantités selon le ratio n/p sur UKBiobank. 110

6.1 Table d'estimation d'héritabilité sur les données de cardiomyopathie. . 137
6.2 Table sur l'influence des différents filtres pour l'estimation d'héritabilité sur les données de cardiomyopathie dilatée. 139
6.3 Table d'influence de δ sur les estimations de PCGC 144

Introduction

Pouvoir quantifier l'influence de la génétique sur un trait biologique, identifier les gènes associés à ce trait et pouvoir réaliser des prédictions de ce trait à partir du génome sont des problèmes d'intérêts multiples avec de nombreuses applications allant de la médecine personnalisée [Jain, 2002, Hamburg and Collins, 2010, Cirillo and Valencia, 2019] à l'industrie agro-alimentaire [Heffner et al., 2009, Goddard, 2009, Hayes et al., 2009]. Dans ce manuscrit de thèse nous proposons un travail qui tente de quantifier la capacité à prédire un trait humain à partir de données génétiques en population. Nous commencerons par présenter le cadre des données que nous utiliserons, présenterons une quantité nommée héritabilité qui sera une pierre angulaire de nos travaux et enfin nous présenterons rapidement le cadre des prédictions en génétique.

Un petit peu de contexte biologique

Les traits biologiques se décomposent en deux familles : les traits mendéliens qui sont déterminés par un seul gène et à l'inverse les traits complexes qui sont le résultat de la combinaison de plusieurs gènes et sur lesquels nous allons nous focaliser. L'étude de ces derniers est la plus complexe car le mécanisme liant le trait et le génotype est souvent inconnu. Le modèle historique et communément adopté comme point de départ pour modéliser les traits quantitatifs est un modèle polygénique sans interaction gène-environnement

$$y = g + f + e$$

où y est un phénotype quantitatif, f un terme confondant, g un terme génétique et e un terme dit d'environnement. Le terme g sera fonction d'un génotype diploïde de p variants et d'un vecteur d'effets génétique $u \in \mathbb{R}^p$.

Les études d'association ont pour objectif de repérer les liens entre des marqueurs génétiques d'intérêt et un phénotype (qualitatif ou quantitatif). La stratégie d'étude

d'association la plus utilisée consiste à tester l'association statistique au sein d'une population entre le phénotype et un très grand nombre de variants situés partout sur le génome : on parle alors d'étude d'association génome entier ou en anglais Genome-Wide Association Studies [Bush and Moore, 2012, Hirschhorn and Daly, 2005, McCarthy et al., 2008]. L'association entre chacun des variants et le phénotype est testée indépendemment, puis une correction pour les tests multiples est appliquée pour prendre en compte le grand nombre de tests réalisés. Ces études ont permis la découverte d'un grand nombre de variants associés à différents traits complexes. Notons toutefois que ces études ne permettent pas de conclure sur la causalité d'un gène pour un trait : elles permettent seulement de mettre en évidence des variants corrélés avec le ou les variants causaux (on parlera alors de variants en déséquilibre de liaison abrégé DL ou en anglais linkage desequilibrium abrégé LD).

Une approche permise par les avancées technologiques récentes est de travailler sur des données de populations, composées d'individus non-apparentés issus d'une (ou plusieurs) population. L'avantage de cette approche est que les cohortes sont beaucoup moins contraignantes à former que pour les études familiales, même si en pratique ces études d'association peuvent présenter des difficultés (telles que la stratification de population qui peut être résumée par des différences de fréquences alléliques entre les populations).

L'héritabilité

Deux définitions de l'héritabilité

Une quantité très utile pour quantifier l'influence de la génétique sur un trait est l'héritabilité, définie comme la proportion de variance phénotypique due à la génétique. Deux définitions de l'héritabilité sont souvent utilisées : l'héritabilité au sens large H^2 qui mesure la contribution globale du génome, et l'héritabilité au sens faible h^2 (également appelée héritabilité additive) qui est définie comme la proportion de variance phénotypique expliquée par les effets additifs des variants. Les calculs d'héritabilité furent formalisés dans les travaux de Fisher [1919]. Ce dernier proposa d'utiliser le modèle polygénique avec ces deux hypothèses supplémentaires :

1. g défini comme la somme de termes génétiques indépendants et
2. e supposé distribué selon une loi normale.

Ces premières estimations d'héritabilité ont été réalisées sur des données familiales, en particulier des études dites de jumeaux pour prendre en compte l'environnement partagé dans une famille. Dans ce contexte, l'héritabilité est une fonction de la corrélation entre le phénotype de deux apparentés. On calculera h^2 si on suppose l'additivité des effets dans le modèle et H^2 si on prend en compte des effets de dominance et d'épistasie (i.e d'interactions entre variants).

L'héritabilité manquante

L'explosion en popularité des GWAS a permis de détecter un grand nombre de variants associés à divers traits biologiques. Ainsi une idée naturelle pour estimer l'héritabilité d'un trait serait d'estimer les effets génétiques avec un modèle linéaire multivarié entre le phénotype d'intérêt et ces variants associés. Malheureusement ces variants n'expliquent qu'une faible partie de la variabilité du trait et l'héritabilité est fortement sous-estimée, donnant lieu au concept d'héritabilité manquante (ou missing heritability en anglais) [Manolio et al., 2009]. Plusieurs hypothèses ont été proposées pour expliquer cette héritabilité manquante : l'influence de variants rares indétectables en GWAS (par manque de puissance statistique), la présence d'interactions gène-environnement ou encore une multitude de variants avec un effet trop faible pour être détecté par les GWAS. Génin [2019] propose une revue de la littérature pour ces différentes explications.

L'estimation d'héritabilité pour le cas quantitatif

Yang et al. ont proposé d'utiliser des données du génome entier d'individus non-apparentés pour estimer cette héritabilité [Yang et al., 2010] à l'aide d'un modèle à effets aléatoires (nous parlerons alors de méthodologie basée sur le REstricted Maximum Likelihood (abrégé REML), nous y reviendrons plus en détail dans le chapitre 4). Un argument pour utiliser ces données de population dans l'estimation d'héritabilité est l'absence d'environnement partagé puisque les individus sont supposés être non-apparentés. La taille est un trait connu pour être fortement influencé par la génétique dont l'héritabilité estimée sur des données familiales est d'environ 80% [Visscher, 2008], tandis que l'héritabilité estimée en utilisant les variants détectés en GWAS est de moins de 5% [Diabetes Genetics Initiative et al., 2008]. Yang et al. estiment avec leur approche l'héritabilité de la taille à environ 45%, expliquant ainsi plus de 50% de l'héritabilité manquante.

D'autres méthodes utilisant des données de génomes entiers ont été proposées pour continuer à réduire l'héritabilité manquante. Cette littérature est extrêmement riche donc cet aperçu ne sera pas exhaustif.

Speed et al. [2012] proposent une alternative pour intégrer le déséquilibre de liaison entre variants causaux et génotypés en développant un système de poids sur les variants. Ces poids sont fonction de la corrélation avec les autres variants génotypés et permettent d'éviter une répétition de l'information qui apparaît quand un variant causal est en déséquilibre de liaison avec plusieurs variants génotypés.

Bulik-Sullivan et al. [2015] proposent une méthode d'estimation d'héritabilité basée sur la régression des statistiques de test (pour l'association entre la réponse et chaque variant) sur une quantité nommée le LD score (définie comme mesurant la quantité de variation génétique capturée par un variant). La manipulation de statistiques de test est très facile et donc cette méthode permet des calculs d'estimation très rapides.

Hou et al. [2019] proposent également un estimateur de l'héritabilité basé sur des statistiques de test et donc mieux adapté aux "Biobanks" (i.e. aux études avec un très grand nombre d'individus) en limitant au maximum les hypothèses sur le vecteur d'effet. Notons également que les auteurs proposent un aperçu de la littérature pour les méthodes dérivées des deux méthodes que nous venons de décrire dans le paragraphe ci-dessus.

Nous citerons également Chen [2014] qui ont proposé une extension de la régression de Haseman and Elston Haseman and Elston [1972], une méthode permettant initialement d'estimer les composantes de variance pour les études de liaisons. Chen propose d'étendre cette régression aux données de population.

Et pour le cas qualitatif

Pour étudier l'héritabilité d'un trait binaire, une approche très utilisée consiste à étudier un phénotype quantitatif sous-jacent (appelé liabilité ou liability en anglais) : un individu avec une liabilité au dessus d'un certain seuil est considéré comme un cas et inversement un individu sous ce seuil est considéré comme un témoin. L'héritabilité est alors calculée à l'échelle de la liabilité. Nous reviendrons sur ce modèle de la liabilité dans le chapitre 6.

En utilisant le modèle de la liabilité, Lee et al. [2011] ont proposé d'estimer l'héritabilité à l'échelle de la liabilité à partir d'une estimation d'héritabilité sur le phénotype binaire. [Golan et al., 2014] ont eux proposé une extension de la régression de Haseman

and Elston pour le cas binaire. Nous reviendrons en détail sur ces méthodes dans le chapitre 6. Notons également que l'approche basée sur le LD score Bulik-Sullivan et al. [2015] fonctionne pour le cas qualitatif.

Prédiction et génétique

Nous avons trouvé les premières traces de l'intérêt pour la prédiction d'effets génétiques en utilisant des données de génotypage dans Meuwissen et al. [2001]. Selon les articles, cette thématique est appelée *sélection génomique* [Rabier et al., 2016], *prédiction génomique* [Karaman et al., 2016] ou encore (plus rarement) *précision de prédiction* [De los Campos et al., 2013]. De nombreuses approches existent pour la sélection génomique et de nombreuses méthodes sont basées sur les BLUP [Speed and Balding, 2014] (nous présenterons ces derniers dans le chapitre 2). Parmi elles se trouvent les *Genomic Best Linear Unbiased Predictor* (GBLUP), une méthode utilisant des BLUP sur les variants que l'on a génotypés (et non les variants causaux que l'on ne connaît pas) et qui a gagné en popularité avec l'explosion de la quantité de données génétiques disponibles. Sur des données animales ou végétales, la sélection génomique a donné de bons résultats (par exemple pour leur application sur les vaches laitières [Hayes et al., 2009]). En particulier les GBLUP se sont montrés plus fiables que les méthodes basées sur les pédigrés [Clark et al., 2012].

La sélection génomique donne de très bons résultats en agroalimentaire mais semble mal s'exporter dans le cadre de la génétique humaine Dandine-Roulland et al. [2016], De los Campos et al. [2013]. Contrairement au monde de l'agroalimentaire, il n'est pas possible de contrôler la diversité génétique et l'environnement en génomique humaine, et les capacités prédictives s'en ressentent. Nous reviendrons plus en détail sur ces points dans le chapitre 5.

Déroulement du manuscrit

Les deux premiers chapitres nous permettront de poser les outils dont nous aurons besoin pour la suite de la thèse. Nous commencerons par brièvement définir le cadre de l'apprentissage statistique et définirons ses notions les plus importantes dans le chapitre 1. Nous nous focaliserons dans le chapitre suivant sur une méthode d'apprentissage statistique appelée régression ridge.

Dans le chapitre 3 nous présenterons nos travaux sur la Generalized Cross Validation, une quantité liée à la paramétrisation de la régression ridge. Nous montrons que cette quantité n'est pas toujours fonctionnelle lorsque le nombre de variables est supérieur au nombre d'observations et proposerons des solutions pour la réparer.

Dans le chapitre 4 nous proposerons de voir l'héritabilité comme un problème de prédiction avec la régression ridge sous plusieurs angles. Pour cela nous nous appuierons sur des résultats sur simulations et sur données réelles. Nous montrerons qu'il est possible de réaliser des estimations satisfaisantes d'héritabilité avec la régression ridge.

Dans le chapitre 5 nous développerons une approximation du comportement de la précision de prédiction pour la régression ridge, encore une fois à l'aide de simulations et de données réelles. Cette section nous permettra d'appuyer le fait qu'une héritabilité élevée n'est pas suffisante pour réaliser de bonnes prédictions.

Nous présenterons ensuite des résultats sur l'estimation d'héritabilité pour des phénotypes binaires. Nous y présenterons des explications pour les biais observés sur des méthodologies de référence.

Chapitre 1

Théorie de l'apprentissage

Dans ce chapitre, nous présenterons rapidement des notions de base de l'apprentissage statistique.

1.1 Convention de notations

Nous présentons ici les notations que nous adopterons pour la suite de ce manuscrit. Établissons tout d'abord que n représentera pour nous le nombre d'observations et p le nombre de variables.

De manière générale les variables aléatoires seront écrites en italique majuscule X. Les réalisations de cette variables s'écriront en italique minuscule, ainsi x_i représente la i-ème réalisation de la variable X (notons que x_i peut être un scalaire ou un vecteur). Nous écrirons les vecteurs colonnes en minuscule italique, sauf pour ceux de taille n qui seront en minuscule gras. Les matrices seront toujours écrites en gras. Prenons par exemple une matrice de génotypes $\mathbf{M} \in \mathcal{M}(\mathbb{R}) = [m_1, ..., m_n]^T = [\mathbf{m}_1, ..., \mathbf{m}_p]$: m_i représente le génotype d'un individu i et $\mathbf{m}_j$ l'ensemble des réalisations pour le variant j.

De manière générale nous appellerons Y nos variables réponse. La notation $\mathbf{M}$ sera associée aux matrices de génotypes mais nous préférerons souvent travailler avec Z qui correspond aux matrices de génotypes normalisés. $\mathbf{X}$ sera utilisé pour les matrices de données non-génétiques. Enfin la lettre E sera associée au bruit statistique.

1.2 Notion d'apprentissage

1.2.1 Terminologie de l'apprentissage

L'objectif de l'apprentissage (aussi appelé Machine Learning) est de permettre la prédiction d'une réponse Y (ou sortie) à partir d'un ensemble de variables discrètes ou continues Z (aussi appelées prédicteurs ou variables d'entrée). Par exemple nous souhaiterons prédire la taille d'un individu avec des prédicteurs tels que son sexe, son âge ou son génotype. On parle de régression si Y est une variable quantitative et de classification si Y est une variable qualitative.

L'objectif de l'apprentissage est la construction d'un apprenant f, une fonction ayant pour arguments des prédicteurs et rendant une prédiction de la réponse $\hat{Y} = f(Z)$. Notons que très souvent les prédicteurs à notre disposition n'expliquent pas intégralement la réponse et ne permettent donc pas des prédictions parfaites i.e. $\hat{Y} \neq Y$. De manière générale on a $Y = f(Z) + E$ une modélisation avec $\mathbb{E}[E] = 0$ et E indépendant de Z. E représente le bruit statistique inexplicable par nos données.

Les questions qui se posent naturellement sont comment construire un apprenant et comment choisir un bon apprenant ? Pour la construction de l'apprenant, nous disposons d'un ensemble de réalisations de Y et X noté $\mathcal{A} = \{(z_1, y_1), ..., (z_n, y_n)\}$ et appelé ensemble d'apprentissage. Pour juger de la qualité d'un apprenant, nous comparerons les réponses Y avec les prédictions $\hat{Y}$ en utilisant une fonction de coût $\mathcal{L}$.

Comme $\mathcal{A}$ est un ensemble fini, il nous sera impossible d'estimer f exactement. En pratique on approchera f avec un estimateur $\hat{f}$ qui est une fonction de $\mathcal{A}$. Notons que nous nous autoriserons à noter $\hat{Y} = \hat{f}(Z)$.

1.2.2 Perte d'erreur quadratique et fonction de régression

Une fonction de coût très populaire et pratique en régression est l'erreur quadratique (ou risque d'erreur quadratique) $\mathcal{L}(Y, f(Z)) = (Y - f(Z))^2$. Nous allons alors chercher f qui minimise l'erreur de prédiction attendue

$$\mathrm{EPE}(f) = \mathbb{E}_{Z,Y}[\mathcal{L}(Y, f(Z))] = \mathbb{E}_{Z,Y}\left[(Y - f(Z))^2\right] \tag{1.1}$$

$$= \mathbb{E}_Z \mathbb{E}_{Y|Z}\left[(Y - f(Z))^2\right]. \tag{1.2}$$

Minimisons cette erreur conditionnellement à Z et cherchons

$$f(z) = \arg\min_q \mathbb{E}_{Y|Z}\left[(Y-q)^2 \,|Z=z\right]. \tag{1.3}$$

En dérivant simplement par q, nous trouvons

$$f(z) = \mathbb{E}_{Y|Z}\left[Y|Z=z\right]. \tag{1.4}$$

Ainsi pour l'erreur de prédiction attendue, la meilleure prédiction de $Y|Z=z$ est la moyenne conditionnelle. Cette solution $f(z)$ est aussi appelée fonction de régression. Ce résultat est particulièrement important car il nous donne une expression pour l'apprenant.

1.3 Un premier exemple de fonction de régression avec le modèle linéaire

Les modèles linéaires sont un ensemble de modèles simples mais très étudiés et utilisés en apprentissage supervisé. Ces modèles supposent une fonction de régression qui est une combinaison linéaire des variables d'entrées. Pour $u = [u_0, u_1, ..., u_p]$ un vecteur de coefficients, la fonction de régression est de la forme

$$f\left(Z\right) = u_0 + \sum_{j=1}^{p} u_j Z_j. \tag{1.5}$$

Le coefficient u_0, appelé intercept, correspond à la valeur $f(Z = 0_p)$ et est donc indépendant de Z. Ce vecteur u est inconnu donc pour pouvoir faire des prédictions nous allons devoir l'estimer à partir de notre ensemble d'apprentissage $\mathcal{A}$.

Une méthode très classique est la méthode des moindres carrés qui vise à choisir u minimisant la somme des carrés des résidus (que nous abrégeons en RSS pour *Residual Sum of Squares*)

$$\begin{aligned} \mathrm{RSS}\left(u\right) &= \sum_{i=1}^{n} \left(y_i - f(z_i)\right)^2 & (1.6)\\ &= \sum_{i=1}^{n} \left(y_i - u_0 - \sum_{j=1}^{p} z_{ij} u_j\right)^2. & (1.7) \end{aligned}$$

Notons que RSS ne force aucune hypothèse sur le modèle : il se contente de mesurer à quel point les données suivent un modèle linéaire.

La forme de l'estimateur de u apparaît clairement si on travaille avec les expressions matricielles. Notons $\mathbf{Z} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ une matrice dont les lignes correspondent à une entrée et dont le premier élément vaut 1 (pour l'intercept). On peut alors réécrire 1.7

$$\hat{u} = \arg\min_{u} \mathrm{RSS}(u) = \|\mathbf{y} - \mathbf{Z}u\|_2^2 . \tag{1.8}$$

En dérivant 1.8 par u et en supposant que $\mathbf{Z}$ soit une matrice de rang $p+1$, nous obtenons

$$\hat{u} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}. \tag{1.9}$$

En notant $z_{te} \in \mathbb{R}^{p+1}$ une nouvelle entrée, nous pouvons utiliser l'expression 1.9 pour obtenir une nouvelle prédiction

$$\hat{y}_{te} = z_{te}^T\hat{u} = z_{te}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}. \tag{1.10}$$

Introduisons le concept de modélisation linéaire additive : supposons que le modèle 1.5 soit vrai i.e. que l'espérance de Y conditionnellement à Z soit une fonction linéaire des composantes de Z. En ajoutant un bruit gaussien additif, nous avons la modélisation

$$Y = u_0 + \sum_{j=1}^{p} u_j Z_j + e \text{ avec } e \sim \mathcal{N}\left(0, \sigma^2\right). \tag{1.11}$$

Cette modélisation est dite linéaire et additive car c'est une combinaison linéaire des prédicteurs. Cette modélisation est simpliste mais extrêmement utile car intuitive : chaque prédicteur apporte sa pierre à la réponse, il n'y a pas d'interaction entre les prédicteurs et ceux-ci sont autorisés à avoir des effets différents. Sous cette modélisation nous voyons que

$$\hat{u} \sim \mathcal{N}(u,\ \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}). \tag{1.12}$$

Remarquons que $\mathbb{E}(\hat{u}) = u$: notre estimateur est dit sans biais.

1.4 Concept de sur-apprentissage et un exemple avec l'estimateur des moindres carrés

Une fois que nous avons choisi un modèle, nous allons souhaiter quantifier sa validité i.e. notre capacité à faire de bonnes prédictions avec. Pour quantifier nous pouvons utiliser une fonction de coût mais sur quelles données allons nous faire des prédictions ?

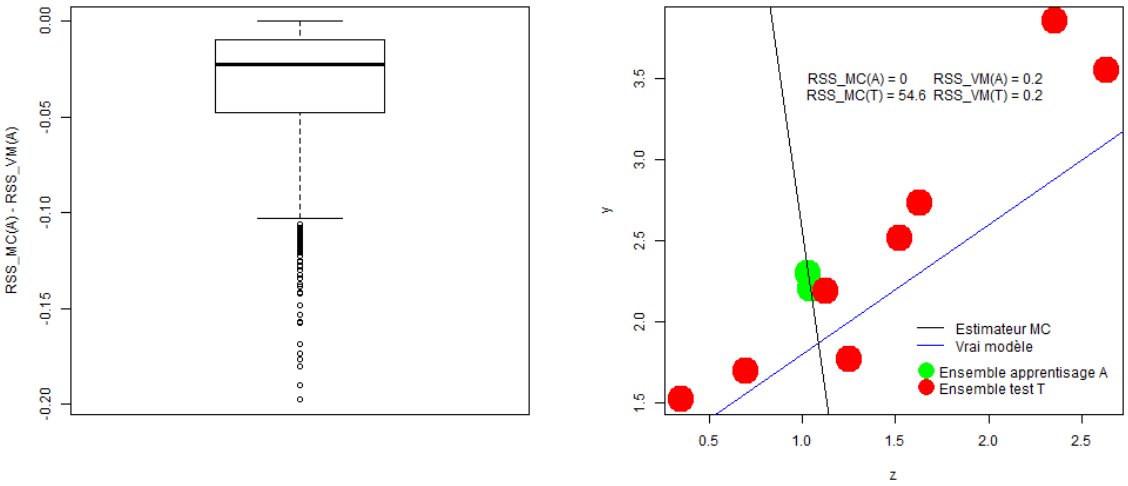
Jusqu'à présent toutes les données dont nous disposons ont été placées dans l'ensemble d'apprentissage (avec lequel nous approchons un modèle par un estimateur). Nous pourrions être tentés d'évaluer l'estimateur sur $\mathcal{A}$ mais nous nous heurterions à la hantise de l'apprentissage statistique : le surapprentissage.

On parle de surapprentissage (ou overfitting) quand l'estimateur s'adapte trop bien à l'ensemble d'apprentissage en dépit du vrai modèle. Pour surveiller ce surapprentissage, une bonne pratique est d'avoir un ensemble de test $\mathcal{T}$. Cet ensemble servira à évaluer la qualité des prédictions obtenues avec notre estimateur, et ne devra donc surtout pas être utilisé pendant la phase de construction de l'estimateur. D'un point de vue formel, une bonne manière de repérer le surapprentissage est de comparer le signe de la différence entre erreur sur $\mathcal{T}$ et l'erreur sur $\mathbb{A}$. Si cette différence est positive, alors il y surapprentissage.

Notons toutefois qu'avoir un ensemble de test ne supprime pas le surapprentissage, il permet seulement de d'estimer de manière non biaisée la qualité des prédictions.

Le surapprentissage s'exprime sous diverses formes et peut être montré de façon très visuelle avec le modèle linéaire et l'estimateur des moindres carrés (voir figure 1.1). Montrons d'abord que l'erreur calculée sur $\mathcal{A}$ est trop optimiste à l'aide de simulations. Nous allons simuler $n = 10$ réponses $y = 1+0.8\times z+e$ avec $z \sim \mathcal{U}(0,3)$ et $e \sim \mathcal{N}(0,\sqrt{3})$. Nous calculerons alors l'estimateur des moindres carrés sur toutes les données que nous avons simulées. Nous calculerons ensuite le RSS pour l'estimateur des moindres carrés RSS_{MC} et pour le vrai modèle RSS_{VM} sur ces mêmes données pour estimer la qualité de la prédiction. Cette procédure est répétée 1000 fois. La figure 1.1a montre un boxplot de la quantité $\mathrm{RSS}_{MC} - \mathrm{RSS}_{VM}$. Nous voyons que l'erreur associée au modèle estimé est plus faible que celle du vrai modèle. Cela montre que l'erreur de prédiction est sous-estimée si nous n'utilisons pas d'ensemble de test.

Dans la figure 1.1b nous allons simuler un ensemble de données comme décrit dans le paragraphe précédent. Nous utilisons maintenant un ensemble de test (en rouge)



(a) Différence entre l'erreur de l'estimateur des moindres carrés et le vrai modèle sur $\mathcal{A}$.

(b) Exemple de surapprentissage avec l'estimateur des moindres carrés.

FIGURE 1.1 – Deux mises en évidence du phénomène de surapprentissage avec l'estimateur des moindres carrés. À gauche on montre que l'erreur de prédiction de l'estimateur calculée sur $\mathcal{A}$ est trop optimiste. À droite on a un exemple évident de surapprentissage car $\mathcal{A}$ est trop petit.

mais nous n'avons que deux de points dans l'ensemble d'apprentissage (en vert). Nous voyons bien ici bien l'importance d'utiliser un ensemble de test : pour les moindres carrés le RSS sur $\mathcal{A}$ est nul mais celui sur $\mathcal{T}$ ne l'est pas du tout. Il y a bien du surapprentissage et l'échantillon de test le révèle complètement. Dans cette exemple, il est très clair que le surapprentissage vient du fait que $\mathcal{A}$ ne contient pas assez de points pour bien apprendre le modèle (ou dit autrement il y a trop de paramètres par rapport à la taille $\mathcal{A}$). Ce sera bien souvent l'explication du surapprentissage.

1.5 Concept d'erreur

Jusqu'à présent nous sommes restés assez évasifs sur le concept d'erreur : nous nous sommes contentés d'utiliser le RSS pour quantifier la qualité d'une régression linéaire. Nous allons ici préciser un peu les choses. Pour un estimateur $\hat{f}$ et une fonction de coût $\mathcal{L}$, nous définissons l'erreur de test comme

$$\mathrm{Err}_{\mathcal{A}} = \mathbb{E}_{(Y,X)\in\mathcal{T}}\left[\mathcal{L}(Y, \hat{f}(X))|\mathcal{A}\right]. \tag{1.13}$$

Elle correspond à l'espérance de l'erreur de prédiction sur un ensemble de test indépendant et avec un ensemble d'apprentissage fixé. Une quantité associée est l'erreur

de test attendue

$$\text{Err} = \mathbb{E}_{\mathcal{A}}\text{Err}_{\mathcal{A}} = \mathbb{E}\left[\mathcal{L}(Y, \hat{f}(X))\right]. \tag{1.14}$$

Pour cette quantité l'ensemble de test ET l'ensemble d'apprentissage sont considérés aléatoires. Notre objectif sera d'estimer $\text{Err}_{\mathcal{A}}$ car nous ne disposerons que d'une quantité limitée de données et donc d'un ensemble d'apprentissage fixé. En pratique nous nous intéresserons souvent à Err car plus facile à estimer (avec des méthodes telles que la validation croisée, nous y reviendrons plus en détail plus tard).

1.6 Sélection et évaluation de modèle

Dans cette section nous présentons le concept de sélection et d'évaluation de modèle. Pour cela nous allons avoir besoin de définir le principe de complexité.

La complexité d'un modèle peut être vue comme la force du lien entre la réponse et les paramètres [Höge, 2016]. La complexité exprime aussi la difficulté du modèle à être appris : plus le modèle est complexe, plus son estimation sera compliquée à réaliser. Dans le cas de l'estimation du modèle linéaire par la régression des moindres carrés, le nombre de variables intégrées au modèle est une mesure de la complexité. Nous reviendrons sur d'autres mesures de complexité un peu plus loin dans ce chapitre.

Le concept de sélection de modèle peut se résumer par la volonté de choisir le meilleur modèle au sens de la prédiction parmi toute une famille de modèles. Supposons que l'on dispose d'un vecteur de réponse et de 10 variables explicatives associées : la recherche de la régression linéaire univariée minimisant une erreur quadratique est un exemple de sélection de modèle.

La sélection de modèle est une étape nécessaire pour des approches de modélisation contrôlant la complexité via un hyperparamètre (que nous appelons ici λ). Pour ces modèles, nous allons donc devoir choisir λ (et donc la complexité optimale) qui minimise l'erreur de test attendue. Il y aura donc trois étapes distinctes.

1. L'étape de création des modèles : à l'aide des données, nous créons une famille de modèles plus ou moins compliqués (i.e. complexes), ce qui en pratique correspond à tester une gamme de valeurs du paramètre λ.
2. L'étape sélection de modèle : nous allons tester chacun des modèles et sélectionner le meilleur possible i.e. le modèle permettant la meilleure prédiction possible.

3. L'étape évaluation de modèle : une fois que nous avons choisi le modèle optimal pour notre problème, nous allons estimer son pouvoir prédictif avec l'erreur de test sur un ensemble $\mathcal{T}$ indépendant.

Pour l'étape évaluation de modèle, nous avons vu qu'un ensemble de test indépendant $\mathcal{T}$ était nécessaire. La même philosophie est à appliquer sur l'étape de sélection pour éviter des biais. Idéalement (si nous avons suffisamment de données) nous les découperons en trois ensembles indépendants : un ensemble d'apprentissage $\mathcal{A}$ sur lequel nous estimerons nos modèles, un ensemble de validation $\mathcal{V}$ avec lequel nous choisirons le modèle minimisant l'erreur de test et enfin un ensemble de test $\mathcal{T}$ sur lequel nous estimerons $\text{Err}_{\mathcal{A}}$. Si nous manquons de données, il est possible d'utiliser des approches pour compenser l'étape de sélection, nous en reparlerons rapidement au prochain chapitre.

Mesures de la complexité

Nous avons défini la complexité mais nous sommes restés assez vagues sur sa définition précise. Les degrés de liberté sont une manière de quantifier cette complexité. Ils sont souvent définis comme le nombre d'observations moins le nombre de paramètres à estimer. Bien que intuitive, cette définition ne s'exporte pas très bien dans le cas $n < p$. De plus elle ne prend pas en compte les potentiels hyperparamètres des modèles.

Soit un vecteur de prédiction $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_n)$ calculé sur les données de $\mathcal{A}$, Trevor Hastie et al. définissent les dégrés de liberté effectifs (en anglais effective degrees of freedom) comme

$$\text{df}(\hat{\mathbf{y}}) = \frac{1}{\sigma^2}\text{tr}(\text{cov}(\mathbf{y}, \hat{\mathbf{y}})). \tag{1.15}$$

Cette définition est bien pratique car elle s'étend complètement à la notion d'hyperparamètres : à chaque valeur d'hyperparamètres λ est associée une valeur des degrés de libertés $\text{df}(\hat{\mathbf{y}}(\lambda))$.

La décomposition Biais-variance

Si nous utilisons la perte d'erreur quadratique comme fonction de coût, l'erreur de test attendue avec individu de test fixé en régression peut être découpée de la façon suivante (preuve en annexe) :

$$\mathbb{E}_{X=x_0}\left[(Y-\hat{f}(x_0))^2\right] = \underbrace{\text{var}(Y)}_{\text{erreur irréductible}} + \underbrace{\text{var}(\hat{f}(x_0))}_{\text{variance estimateur}} + \underbrace{\left(\mathbb{E}_{X=x_0}\left[\hat{f}(x_0)\right] - \mathbb{E}_{X=x_0}[Y]\right)^2}_{\text{biais}}. \tag{1.16}$$

Le premier terme est l'erreur irréductible de notre problème qui ne peut être évitée quel que soit notre apprenant. Le deuxième terme représente la variance de notre estimateur. Enfin le dernier terme représente le carré du biais de l'estimateur i.e. le carré de la différence entre l'espérance de l'estimateur et le vrai modèle. Le biais et la variance sont des fonctions de la complexité : le terme de biais va diminuer quand la complexité augmente et à l'inverse le terme de variance va augmenter avec la complexité.

Le découpage 1.16 est appelé la décomposition biais-variance et est particulièrement utile dans l'étape de sélection de modèle. Comme les termes de biais et de variance sont monotones, le minimum de l'erreur de prédiction correspondra à la complexité proposant le meilleur compromis entre biais et variance.

Le dilemme biais-variance est souvent présenté pour l'erreur quadratique de l'estimateur. Dans ce cas la décomposition devient

$$\mathbb{E}\left[(\beta-\hat{\beta})^2\right] = \underbrace{\text{var}(\hat{\beta})}_{\text{variance estimateur}} + \underbrace{\left(\mathbb{E}\left[\hat{\beta}\right]-\beta\right)^2}_{\text{biais}} \tag{1.17}$$

Notons qu'ici nous utilisons l'abus de notation $x^2 = x^T x$ pour les vecteurs et le terme $\text{var}(x)$ se réfère à la variance des coefficients de x et non à la matrice de covariance de x.

Régression pénalisée

Si nous considérons le critère des moindres carrés défini en 1.6 pour un apprenant f quelconque, nous voyons que le critère peut être minimisé par une infinité d'apprenants (tous ceux vérifiant $y_i = f(z_i)\ \forall i \in [\![1, n]\!]$). Pour le cas des modélisations linéaires, ce phénomène apparaît pour le cas $n < p$ et donc nous ne pourrons nous contenter de minimiser un critère des moindres carrés (abrégé MC) pour estimer le modèle.

Pour pouvoir utiliser un modèle linéaire dans le cas $n < p$, une famille d'approches très utilisée sont les régressions pénalisées : nous allons chercher à minimiser un critère

de la forme suivante

$$\mathrm{PRSS}\,(f, \lambda) = \mathrm{RSS}(f) + \lambda J(f). \tag{1.18}$$

Le terme $J(f)$ est un terme dit de pénalité et donne un scalaire pour chaque f. Nous choisirons J croissant en la complexité de f, i.e. l'inverse du terme RSS : la pénalité a pour objectif de "contenir" la courbe i.e. d'éviter de sélectionner un modèle très complexe adhérant très bien à l'ensemble d'apprentissage et se généralisant très mal. Nous choisirons alors f comme minimisant le compromis entre les deux critères, i.e. la solution la plus proche de nos données et restant la moins complexe possible. Pour régler ce compromis nous jouerons sur l'hyperparamètre $\lambda \geq 0$ appelé constante de pénalisation. Plus cette constante sera élevée, plus nous pénaliserons l'aspérité de la solution et donc nous aurons tendance à choisir des solutions "simples". Les extrêmes des valeurs de λ illustrent bien ce phénomène : le cas $\lambda = 0$ correspond au cas où nous n'imposons aucune pénalisation à la complexité de notre modèle tandis que le cas $\lambda = +\infty$ impose la solution la moins complexe $f = 0$.

Dans le cas des modèles linéaires, nous pouvons citer comme exemple le LASSO [Tibshirani, 1996] ou la régression ridge que nous allons présenter plus longuement dans la suite de ce manuscrit. En effet, nous verrons (entre autres) dans le chapitre suivant que la régression ridge permet un parallèle avec les modèles à effets aléatoires, une méthode mathématique très utilisée pour l'estimation d'héritabilité.

Chapitre 2

La régression ridge

Le chapitre précédant nous a permis de poser le cadre de l'apprentissage statistique. Nous allons maintenant nous concentrer une méthode d'apprentissage appelée régression ridge. Nous présenterons la méthode, quelques unes de ses propriétés et discuterons de la gestion de son hyperparamètre. Nous présenterons également très brièvement les modèles à effets aléatoires et discuterons de leurs liens avec la régression ridge.

2.1 Définition et origine

La régression d'arête (que l'on anglicisera en régression ridge) a été présentée parallèlement dans différents domaines. Tikhonov la présenta sous le nom de régularisation de Tikhonov comme une méthode permettant la résolution de systèmes linéaires mal posés (i.e. de systèmes n'ayant pas de solution vérifiant les critères d'Hadamard : existance, unicité, et continuité) [Tikhonov, 1965]. Hoerl présenta la méthode sous une approche plus statistique qui nous intéressera plus ici [Hoerl and Kennard, 2000] : l'idée est d'ajouter un petit coefficient au terme à inverser dans l'estimateur des moindres carrés pour contrôler l'instabilité de l'estimateur.

La régression ridge est une régression linéaire pénalisée dont la pénalité est la somme des carrés des coefficients du vecteur d'effets. Supposons que nos données suivent un modèle linéaire

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e} \text{ avec } \mathbf{e} \sim \mathcal{N}\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right). \tag{2.1}$$

Alors l'estimateur de la régression ridge $\hat{u}_R$ est défini comme la solution du critère

$$\arg\min_{u\in\mathbb{R}^p} \mathrm{RSS} + J_{RIDGE} = \arg\min_{u\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{Z}u\|_2^2 + \lambda \|u\|_2^2 . \tag{2.2}$$

Le paramètre de pénalité $\lambda \geq 0$ sanctionne les solutions dont les valeurs de coefficients sont très élevées au profit des solutions dont les coefficients sont plus faibles. Nous dirons alors que la régression ridge "lisse" les solutions. Sous l'hypothèse que l'espérance des effets est nulle, la ridge peut être vue comme une régression linéaire qui contrôle la variance des ses coefficients. Remarquons également que le cas $\lambda = 0$ correspond au cas des moindres carrés ordinaires tandis que pour le cas $\lambda = +\infty$ nous choisirons toujours la solution nulle.

La minimisation du critère 2.2 donne un estimateur avec une forme close et de nombreuses propriétés. Nous allons maintenant présenter quelques unes des ces propriétés.

2.2 Standardisation des données

Il est très fortement conseillé d'appliquer la régression ridge sur un modèle dont la réponse est centrée et la matrice de données est centrée et réduite par colonnes.

Travailler sur des données non-centrées revient à ajouter au modèle un terme constant appelé intercept, qui s'interprète comme la moyenne de la réponse. En regardant le critère 2.2, nous voyons que toutes les variables sont concernées par la pénalisation. Si les données ne sont pas centrées nous allons pénaliser cette moyenne, ce qui n'a pas de sens. Si les données sont centrées le problème se résout car la pénalisation n'affecte que la variance.

Sans normalisation de nos données, les estimations des effets risquent d'être à des échelles complètement différentes. Pour visualiser ce phénomène, imaginons un modèle visant à expliquer la taille d'une personne en fonction de son âge et de son sexe. Selon que l'âge soit en années ou en mois, l'effet estimé sera très différent. La normalisation de la matrice des données permet de s'assurer que tous les effets soient à la même échelle. Or nous avons dit que la pénalisation de la régression ridge pouvait être vue comme un contrôle de variance du vecteur d'effets. Puisque cette pénalisation est la même pour toutes les variables, en l'absence de normalisation des données nous pénalisons autant toutes les variables (même si elles sont à des échelles très différentes). La normalisation

des données (et donc le fait d'avoir les effets à la même échelle) donne une cohérence à la pénalisation unique.

En pratique nous aurons tendance à utiliser la moyenne empirique de notre réponse pour centrer celle-ci. De même nous centrerons et réduirons chaque colonne des données par respectivement sa moyenne et son écart-type empirique. Bien que très courante, ces pratiques peuvent-être discutées (particulièrement dans le cas où $n < p$) et nous reviendrons dessus plus tard.

2.3 Calcul de l'estimateur

Comme la régression linéaire standard, l'estimateur de la régression ridge dispose d'une formule simple. Dérivons selon u le critère 2.2 :

$$\begin{aligned}\partial_u \left(\|\mathbf{y} - \mathbf{Z}u\|_2^2 + \lambda \|u\|_2^2 \right) &= \partial_u \left((\mathbf{y} - \mathbf{Z}u)^T (\mathbf{y} - \mathbf{Z}u) + \lambda u^T u \right) \\ &= 2\mathbf{Z}^T\mathbf{Z}u - 2\mathbf{Z}^T\mathbf{y} + 2\lambda u.\end{aligned}$$

En posant $\hat{u}_R$ le vecteur qui minimise 2.2 nous pouvons écrire

$$\partial_{u=\hat{u}_R} \left(\|\mathbf{y} - \mathbf{Z}u\|_2^2 + \lambda \|u\|_2^2 \right) = 0 \tag{2.3}$$

$$\Rightarrow \hat{u}_R = \left(\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p \right)^{-1} \mathbf{Z}^T\mathbf{y} \tag{2.4}$$

2.4 Propriétés de l'estimateur

2.4.1 Décomposition en valeurs singulières, analyse en composantes principales et degrés de liberté

La décomposition en valeurs singulières est une technique très utilisée en analyse des données [Golub and Kahan, 1965, Golub and Reinsch, 1970]. Nous appellerons décomposition en valeurs singulières (SVD) de $\mathbf{Z}$ la décomposition $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. $\mathbf{U} \in \mathcal{O}(n)$ et $\mathbf{V} \in \mathcal{O}(p)$ décrivent respectivement l'espace des lignes et des colonnes de $\mathbf{Z}$. $\mathbf{D} \in \mathcal{D}_{n,p}(\mathbb{R})$ est une matrice diagonale dont les coefficients diagonaux sont appelés valeurs singulières. Notons que toute matrice $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ admet au maximum $\min(n, p)$

valeurs singulières non nulles.

L'analyse en composantes principales (ACP), également connue sous le nom de transformation de Karhunen–Loève est une méthode permettant de construire un jeu de variables décorrélées décrivant le mieux un jeu de données. Elle trouve ses origines dans Pearson [1901] et fut formalisée par Hotelling [1933].

La SVD et l'ACP sont complètement liées : l'ACP de la matrice $\mathbf{Z}$ correspond à la SVD de la matrice $\mathbf{Z}^T\mathbf{Z}$. En particulier, les valeurs singulières correspondent aux racines carrées des valeurs propres.

En utilisant la SVD de $\mathbf{Z}$ nous pouvons écrire l'estimateur de la régression ridge sous la forme suivante :

$$\hat{u}_R = \mathbf{V}\left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{y}. \tag{2.5}$$

Cette forme est intéressante d'un point de vue calculatoire car elle demande l'inversion d'une matrice diagonale et non pleine comme 2.4, au prix d'une décomposition en valeurs singulières de $\mathbf{Z}$ (qui est également coûteuse en temps de calcul). C'est toutefois très intéressant si nous souhaitons calculer l'estimateur pour plusieurs valeurs du paramètre de pénalisation, car nous n'avons qu'à calculer la SVD de $\mathbf{Z}$ une seule fois. Nous en parlerons plus en détail plus loin dans le manuscrit.

La décomposition en valeurs singulières permet de définir pour la régression ridge une forme particulière des degrés de liberté effectifs. Notons $\hat{\mathbf{y}} = \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{y} = \mathbf{H}_\lambda\mathbf{y}$ le vecteur de prédiction sur l'ensemble d'apprentissage. En reprenant la définition des degrés de liberté effectifs 1.15, la formule d'espérance quadratique et la SVD

$$\begin{aligned}
\mathrm{df}(\hat{\mathbf{y}}) &= \frac{1}{\sigma^2}\mathrm{tr}\left(\mathrm{cov}(\hat{\mathbf{y}}, \mathbf{y})\right) = \frac{1}{\sigma^2}\mathbb{E}_{\mathbf{e}}\mathbf{e}^T\mathbf{H}_\lambda\mathbf{e} && (2.6)\\
&= \frac{1}{\sigma^2}\mathrm{tr}\left(\sigma^2\mathbf{U}\mathbf{D}\left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{D}^T\mathbf{U}^T\right) && (2.7)\\
&= \sum_{k=1}^{\min(n,p)} \frac{d_k^2}{d_k^2 + \lambda}. && (2.8)
\end{aligned}$$

Un intérêt de cette quantité est qu'elle est bornée $0 \leq \mathrm{df}(\hat{\mathbf{y}}) \leq \min(n, p)$. La quantité df augmente avec la complexité du modèle : $\mathrm{df} = \min(n, p)$ correspond au modèle linéaire non pénalisé et $\mathrm{df} = 0$ correspond au modèle nul.

2.4.2 Le dilemme biais-variance pour la régression ridge

En supposant la matrice de données $\mathbf{Z}$ fixe, le dilemme biais-variance décrit en 1.17 pour la régression ridge s'écrit de la manière suivante :

$$\begin{aligned}
\text{var}(\hat{u}_R) &= \mathbb{E}_{\mathbf{y}}\left[(\hat{u}_R - \mathbb{E}_{\mathbf{y}}[\hat{u}_R])^T(\hat{u}_R - \mathbb{E}_{\mathbf{y}}[\hat{u}_R])\right] \\
&= \mathbb{E}_{\mathbf{y}}\left[(\hat{u}_R - \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{Z}u)^T(\hat{u}_R - \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{Z}u)\right] \\
&= \mathbb{E}_{\mathbf{y}}\left[\left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}u)\right)^T\left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}u)\right)\right] \\
&= \text{tr}\left(\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-2}\mathbf{Z}^T\sigma^2\mathbf{I}_n\right) \text{ (espérance d'une forme quadratique)}
\end{aligned}$$

$$\begin{aligned}
\text{biais}^2(\hat{u}_R) &= \left(\mathbb{E}_{\mathbf{y}}[\hat{u}_R] - u\right)^T\left(\mathbb{E}_{\mathbf{y}}[\hat{u}_R] - u\right) \\
&= \left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{Z}u - u\right)^T\left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{Z}u - u\right) \\
&= u^T\left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{Z} - \mathbf{I}_p\right)^2 u \\
&= \lambda^2 u^T\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-2} u
\end{aligned}$$

Ce qui nous donne

$$\text{var}(\hat{u}_R) = \sigma^2 \sum_{k=1}^{\min(n,p)} \frac{d_k^2}{(d_k^2 + \lambda)^2} \tag{2.9}$$

$$\text{biais}^2(\hat{u}_R) = \lambda^2 \sum_{k=1}^{\min(n,p)} \frac{c_k^2}{(d_k^2 + \lambda)^2} \text{ avec } c = \mathbf{V}^T u. \tag{2.10}$$

La variance et le biais sont respectivement des fonctions décroissante et croissante en λ (voir annexe). En pratique la régression ridge propose donc de biaiser l'estimateur des moindres carrés pour baisser sa variance, et l'objectif sera donc de trouver le meilleur compromis entre les deux.

2.4.3 Théorème de l'existence

Une propriété intéressante est que l'estimateur de la régression ridge $\hat{u}_R$ peut toujours faire mieux au sens de la précision que l'estimateur des moindres carrées $\hat{u}$.

$$\exists \lambda > 0,\ \mathrm{MSE}(\hat{u}_R, u) \leq \mathrm{MSE}(\hat{u}, u) \tag{2.11}$$

avec $\mathrm{MSE}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\mathbf{x}^T\mathbf{y}$. Une démonstration est disponible en annexe A.3. Sous réserve de connaître un lambda vérifiant cette propriété, il est donc toujours avantageux d'utiliser la régression ridge. Nous verrons plus tard comment choisir ce paramètre.

2.4.4 Un exemple pratique

Un exemple de décomposition biais variance pour l'estimateur de la ridge est présenté dans la figure 2.1. Dans ce graphe nous avons simulé un modèle additif de dimension $n = 100$ et $p = 10$ puis nous avons calculé le biais et la variance avec les formules décrites en 2.9 et 2.10. En abscisse on utilise les d.d.l.e avec la formule 2.8. La courbe rouge représente la variance et la courbe verte est le biais. On voit que la variance augmente quand le modèle se complexifie et qu'à l'inverse le biais diminue avec la complexité. La courbe en noir est la précision de l'estimateur et présente un minimum vers d.d.l.e $\simeq$ 8,5-9. La courbe en pointillé représente la précision de l'estimateur des moindres carrés. Nous voyons bien qu'il existe des valeurs de λ telles que la précision de la régression ridge soit inférieure à celle des moindres carrés.

2.4.5 La régression ridge en grande dimension

Un estimateur toujours défini

Un intérêt de la régression ridge par rapport aux moindres carrés est qu'elle peut être utilisée dans le cas de la grande dimension i.e $n < p$. En effet si $n < p$ alors

$$0 \in \mathrm{sp}\left(\mathbf{Z}^T\mathbf{Z}\right) \Rightarrow \left|\mathbf{Z}^T\mathbf{Z}\right| = \prod_{i=1}^{n} d_i^2 \times 0^{p-n} = 0.$$

La matrice $\left(\mathbf{Z}^T\mathbf{Z}\right)$ n'est donc pas inversible et l'estimateur des moindres carrés ne peut donc pas être défini. Dans le cas de la régression ridge,

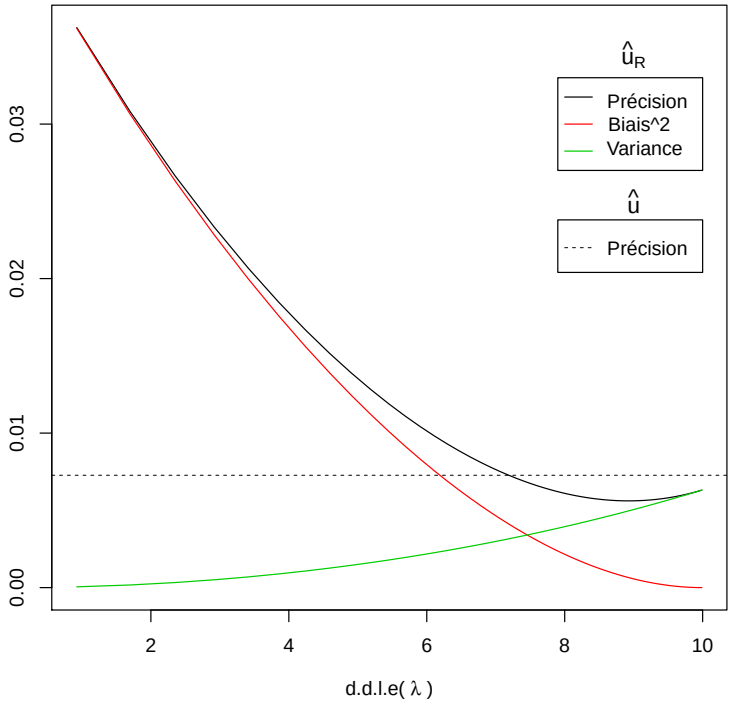


FIGURE 2.1 – Un exemple de la décomposition biais variance de l'estimateur de la régression ridge.

$$\forall \lambda > 0, \ \left|\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p\right| = \prod_{i=1}^{n}(d_i^2 + \lambda) = \lambda^{n-p}\prod_{j=1}^{p}(d_i^2 + \lambda) > 0.$$

L'estimateur de la régression ridge est toujours défini tant que $\lambda > 0$. Dans le cas $n < p$ la ridge se révèle donc un choix intéressant pour une modélisation linéaire qui rend le problème soluble sans sélectionner des variables.

Une forme adaptée à la grande dimension

Dans un cas où p est très grand, le calcul de l'estimateur peut être lourd. En effet l'inversion de la matrice $p \times p$ dans la forme 2.4 va être très compliquée. Nous avons vu qu'utiliser la SVD rend l'inversion plus facile, mais la SVD reste une opération coûteuse. Il existe une forme alternative de l'estimateur de la régression ridge qui évite la manipulation de matrice $p \times p$. En particulier, cette forme sera très intéressante pour le calcul de $\mathbf{H}_\lambda$ en grande dimension.

$$\begin{aligned}
&\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)=\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{I}_n\right)\mathbf{Z}^T\\
\Rightarrow&\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)^{-1}=\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{I}_n\right)^{-1}\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{I}_n\right)\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)^{-1}\\
\Rightarrow&\left(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{Z}^T=\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)^{-1}
\end{aligned}$$

Nous pouvons alors donc alors écrire l'estimateur ridge sous la forme

$$\hat{u}_R=\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)^{-1}\mathbf{y}=\mathbf{V}\mathbf{D}^T\left(\mathbf{D}\mathbf{D}^T+\lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^T\mathbf{y} \tag{2.12}$$

et $\mathbf{H}_\lambda$ sous la forme

$$\mathbf{H}_\lambda=\mathbf{Z}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T+\lambda\mathbf{I}_p\right)^{-1}=\mathbf{U}\mathbf{D}\mathbf{D}^T\left(\mathbf{D}\mathbf{D}^T+\lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^T=\mathbf{U}\boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}\mathbf{U}^T.$$

Sous cette forme de $\mathbf{H}_\lambda$ nous ne manipulons que des matrices $n\times n$. De plus nous ne calculerons pas la SVD de $\mathbf{Z}$ mais plutôt celle de $\mathbf{Z}\mathbf{Z}^T$ (ce qui revient à faire une ACP).

2.5 Choix du paramètre de pénalisation λ

2.5.1 Comment choisir λ ?

Nous avons vu qu'un bon choix de paramètre de pénalisation est important pour avoir un estimateur ridge de qualité. Mais comment choisir ce paramètre ? Dans l'exemple du graphe 2.1 le paramètre optimal λ_{opt} était visible avec la décomposition biais-variance. En pratique nous avons rarement accès au paramètre σ^2 et au vecteur d'effets u et ce n'est donc pas utilisable.

Notre objectif est d'avoir les meilleures prédictions possibles pour l'estimateur ridge i.e de trouver λ_{opt} qui minimise l'erreur de prédiction définie en 1.13 pour un nouveau point (y_0, z_0)

$$\lambda_{opt}=\arg\min_\lambda \mathrm{Err}_{\mathcal{A}}(\lambda)=\arg\min_\lambda \mathbb{E}_{y_0,z_0}\left[\left(y_0-z_0^T\hat{u}_R(\lambda)\right)^2\mid\mathcal{A}\right]$$

avec $z_0\in\mathbb{R}^p$ un vecteur colonne.

Une approche tentante serait de choisir λ à partir des données de l'ensemble d'apprentissage $\mathcal{A}$ i.e choisir le lambda qui minimise l'erreur de prédiction sur $\mathcal{A}$

$$\mathrm{err}_{\mathcal{A}}(\lambda) = \frac{1}{n_{\mathcal{A}}} \sum_{i \in \mathcal{A}} \left(y_i - z_i^T \hat{u}_R(\lambda) \right)^2 \text{ avec } n_{\mathcal{A}} = \mathrm{Card}(\mathcal{A}). \tag{2.13}$$

En pratique cela ne fonctionne pas. En effet $\mathrm{err}_{\mathcal{A}}$ a tendance à sous-estimer $\mathrm{Err}_{\mathcal{A}}$. De plus $\mathrm{err}_{\mathcal{A}}$ favorise toujours le modèle le plus complexe possible et donc choisit $\lambda_{opt} = 0$. Ce phénomène est bien illustré dans la figure 2.2. Dans cet exemple nous avons simulé un jeu de données de taille $n = 600$ et $p = 30$ selon un modèle linéaire. Nous avons gardé 100 individus pour estimer le vecteur d'effet en utilisant la régression ridge et utilisé les 500 autres comme ensemble de validation. Nous avons calculé $\mathrm{err}_{\mathcal{A}}$ (courbe rouge) et estimé $\mathrm{Err}_{\mathcal{A}}$ (courbe bleue) avec le MSE sur l'ensemble de validation. Il apparaît clairement que $\mathrm{err}_{\mathcal{A}}$ est un mauvais estimateur de $\mathrm{Err}_{\mathcal{A}}$.

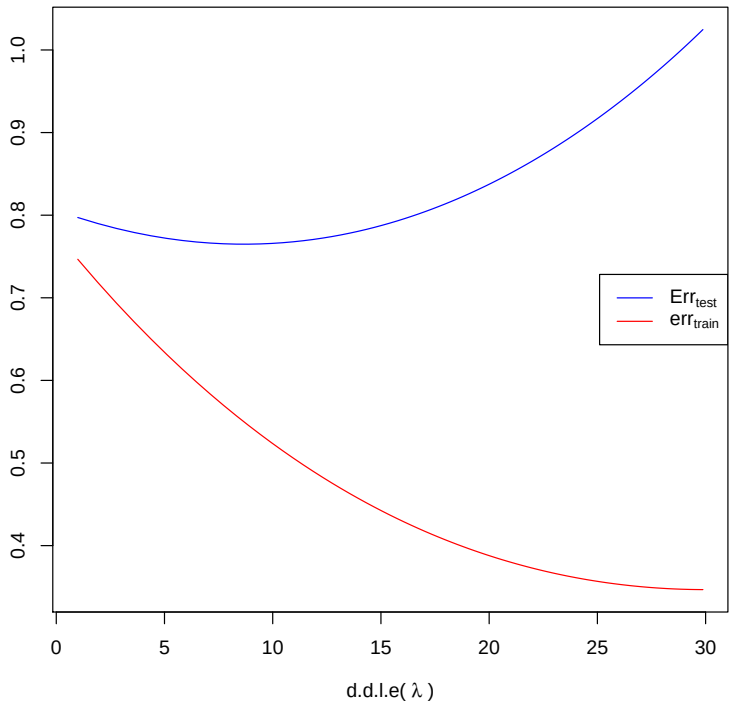


FIGURE 2.2 – Un exemple du comportement de $\mathrm{err}_{\mathcal{A}}$ et $\mathrm{Err}_{\mathcal{T}}$.

Au vu de l'importance d'avoir des données externes pour le choix de λ, une bonne pratique serait de séparer notre jeu de données en 3 parties : un ensemble d'apprentissage $\mathcal{A}$ pour construire l'estimateur, un ensemble de validation $\mathcal{V}$ pour choisir la

complexité optimale (à travers le choix de λ) et un ensemble de test $\mathcal{T}$ pour estimer les capacités prédictives.

Le problème de ce découpage est qu'il est gourmand en données. Il est nécessaire d'avoir une taille d'ensemble d'apprentissage suffisante pour estimer correctement le vecteur d'effet, un faible effectif dans $\mathcal{V}$ peut entraîner un mauvais choix de complexité optimale et enfin l'estimation de la capacité prédictive ne sera pas fiable si on a trop peu d'individus dans $\mathcal{T}$. Ce découpage n'est donc pas toujours applicable en pratique.

Pour palier à ce problème il existe des méthodes pour remplacer l'étape de validation. Une première approche est d'utiliser une formule analytique telle que l'AIC (Aikaike Information Criteria) [Akaike and BN Petrov ; F Csaki, 1973] ou la statistique C_p [Mallows, 1973]. Ces formules permettent de choisir λ à partir de $\mathcal{A}$ et permettent donc de se passer de l'ensemble de validation. Une autre approche est d'utiliser un ré-échantillonnage des données tel que la validation croisée ou le bootstrap. Dans la suite de ce manuscrit nous nous concentrerons sur la validation croisée.

2.5.2 La validation croisée

La validation croisée K fold

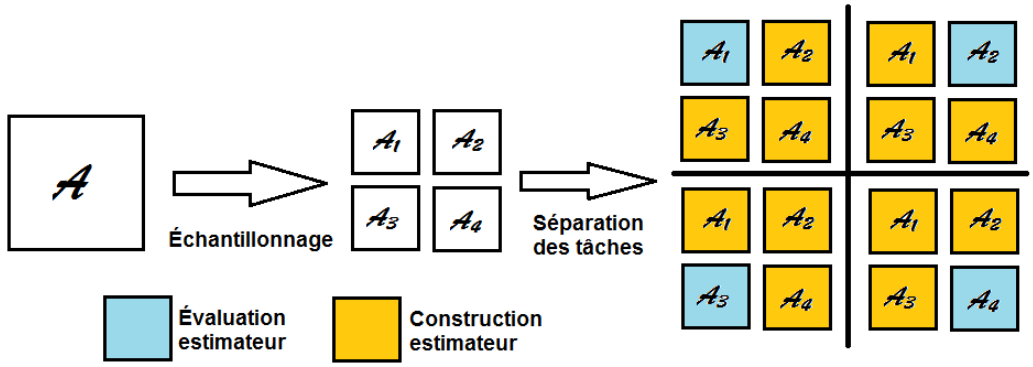


FIGURE 2.3 – Principe de la validation croisée.

La validation croisée vise à fractionner l'ensemble d'apprentissage en sous-ensembles indépendants, puis à construire le modèle sur tous les sous-ensembles sauf un et à l'évaluer sur le sous-ensemble restant, et ceci en changeant plusieurs fois les rôles comme décrit dans le graphe 2.3. Notons $\{\mathcal{A}_1, ..., \mathcal{A}_K\}$ une partition en K sous-ensembles de $\mathcal{A}$ telle que $\mathcal{A} = \bigcup_{l=1}^{K} \mathcal{A}_l$, $\bigcap_{l=1}^{K} \mathcal{A}_l = \emptyset$ et notons également $\mathcal{A}^{-i} = \mathcal{A} \setminus \{\mathcal{A}_l : i \in \mathcal{A}_l\}$ la partition privée du sous-ensemble contenant l'individu i. Nous définissons alors l'erreur de la validation croisée K-fold pour la régression ridge comme

$$\text{err}_{VC}(\lambda) = \frac{1}{n_{\mathcal{A}}} \sum_{i=1}^{n_{\mathcal{A}}} \left(y_i - z_i^T \hat{u}_R^{\mathcal{A}^{-i}}(\lambda) \right)^2, \tag{2.14}$$

avec $z_i \in \mathbb{R}^p$ un vecteur colonne représentant la i-ème ligne de $\mathbf{Z}$ et $\hat{u}_R^{\mathcal{A}^{-i}}$ l'estimateur ridge construit avec $\mathcal{A}^{-i}$. Nous choisirons ensuite

$$\hat{\lambda}_{opt}^{VC} = \arg\min_{\lambda} \text{err}_{VC}(\lambda).$$

Chaque individu est utilisé $K-1$ fois dans la construction d'estimateur et une fois dans l'évaluation du modèle. Très souvent nous prendrons K entre 5 et 15. La validation croisée est une méthode robuste donnant des résultats satisfaisants mais qui peut se révéler parfois coûteuse en temps de calcul.

Notons enfin que la validation croisée estime l'erreur attendue de prédiction 1.14 plutôt que l'erreur de prédiction 1.13 : la validation croisée "construit" plusieurs ensembles de test et d'apprentissage et mimique donc une espérance sur l'ensemble d'apprentissage (voir le chapitre 7 de [Trevor Hastie et al., 2009]).

Un cas particulier : la Leave-One-Out

Un cas particulier de la validation croisée est le cas $K = n$ dans lequel nous effectuons une prédiction sur un individu à partir de tous les autres. Ce cas particulier est appelé validation croisée Laissée pour compte (que nous allons angliciser en Leave-One-Out et abréger en LOO). L'erreur de LOO avec standardisation unique pour tout les individus est définie comme

$$\text{err}^{LOO}(\lambda) = \frac{1}{n_{\mathcal{A}}} \sum_{i=1}^{n_{\mathcal{A}}} \left(y_i - z_i^T \hat{u}_R^{-i}(\lambda) \right)^2 \tag{2.15}$$

avec $\hat{u}_R^{-i} = \mathbf{Z}_{-i} \left((\mathbf{Z}_{-i})^T \mathbf{Z}_{-i} + \lambda \mathbf{I}_p \right)^{-1} (\mathbf{Z}_{-i})^T \mathbf{y}_{-i}$ l'estimateur de la régression ridge construit en excluant la ligne i du vecteur de réponse et de la matrice des données. En utilisant la formule de Sherman-Morrison-Woodbury, Meijer and Goeman [2013] montre que

$$\hat{u}_R^{-i} = \hat{u}_R - \frac{(\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p)^{-1} z_i (y_i - z_i^T \hat{u}_R)}{1 - [h_\lambda]_{ii}} \tag{2.16}$$

avec $[h_\lambda]_{ii}$ le i-ème coefficient diagonal de $\mathbf{H}_\lambda$. Des éléments de la preuve de cette formule sont disponibles en annexe A.4. En injectant 2.16 dans 2.15, nous obtenons

$$\text{err}^{LOO}(\lambda) = \frac{1}{n_{\mathcal{A}}} \sum_{i=1}^{n_{\mathcal{A}}} \left(\frac{y_i - z_i^T \hat{u}_R(\lambda)}{1 - [h_\lambda]_{ii}} \right)^2 \tag{2.17}$$

$$= \frac{1}{n_{\mathcal{A}}} \mathbf{y}^T (\mathbf{I}_n - \mathbf{H}_\lambda)(\text{diag}(\mathbf{I}_n - \mathbf{H}_\lambda))^{-2}(\mathbf{I}_n - \mathbf{H}_\lambda)\mathbf{y}. \tag{2.18}$$

La formule 2.18 est plus rapide à calculer qu'une validation croisée à $K = n_{\mathcal{A}}$ plis (que l'on appellera validation croisée à $K = n_{\mathcal{A}}$ fold) car elle ne nécessite que le calcul de $\mathbf{H}_\lambda$. Notons toutefois qu'il existe une différence subtile au niveau de la normalisation entre l'erreur proposée par Meijer 2.15 et une validation croisée avec $K = n_{\mathcal{A}}$ folds. Dans la version de Meijer $\mathbf{Z}_{-i}$ correspond à la matrice $\mathbf{Z}$ privée de la ligne i et n'est jamais renormalisée. A l'inverse dans une validation croisée avec $K = n_{\mathcal{A}}$ folds, $\mathbf{Z}_{-i}$ sera renormalisée a chaque fois. Cette différence peut paraître anecdotique mais elle peut avoir une grande importance en grande dimension, comme nous allons le voir.

La LOO est l'approche la plus gourmande en calcul pour l'estimation de l'erreur attendue de prédiction parmi la famille des K-fold. Il est donc particulièrement pertinent d'utiliser la LOO quand nous ne disposons que de peu de données. Nous pourrons également souhaiter utiliser la LOO dans le cadre de la grande dimension (i.e. quand $n << p$) où nous souhaiterons utiliser la "meilleure" approximation de l'erreur attendue de prédiction pour "compenser" l'apprentissage difficile. Malheureusement dans ce cadre les temps de calcul de la LOO risquent d'être très longs (par exemple en génétique nous aurons facilement $n > 10\ 000$). Nous allons donc présenter une approche réduisant les temps de calculs pour permettre d'approcher la LOO dans ce contexte de grande dimension.

Une approximation de la Leave-One-Out : la Generalized Cross-Validation

L'erreur de Generalized Cross-Validation (GCV) est une approximation de l'erreur LOO 2.15 proposée dans [Golub et al., 1978]. L'idée est de projeter le modèle dans un espace complexe pour obtenir une matrice $\mathbf{H}_\lambda$ à coefficients diagonaux constants. En combinaison avec la décomposition en valeurs singulières et l'expression 2.17, nous arrivons à une expression très simple :

$$\text{err}^{GCV}(\lambda) = \frac{1}{n_{\mathcal{A}}} \frac{\sum_{k=1}^{n_{\mathcal{A}}} \left(\frac{\lambda}{d_k+\lambda}\right)^2 b_k^2}{\left(\frac{1}{n_{\mathcal{A}}} \sum_{k=1}^{n_{\mathcal{A}}} \frac{\lambda}{d_k+\lambda}\right)^2} \tag{2.19}$$

$$= \frac{1}{n_{\mathcal{A}}} \frac{\|\mathbf{y} - \mathbf{Z}\hat{u}_R(\lambda)\|_2^2}{\left(\frac{1}{n_{\mathcal{A}}} \text{tr}(\mathbf{I}_{n_{\mathcal{A}}} - \mathbf{H}_\lambda)\right)^2} \tag{2.20}$$

avec $\mathbf{b} = \mathbf{U}^T\mathbf{y}$. Il est possible de donner à l'erreur de GCV une forme telle que 2.18 en utilisant l'écriture $\mathbf{H}_\lambda = \mathbf{U}\boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}\mathbf{U}^T$,

$$\text{err}^{GCV}(\lambda) = \frac{1}{n_{\mathcal{A}}} \mathbf{y}^T(\mathbf{I}_{n_{\mathcal{A}}} - \mathbf{H}_\lambda)(\text{tr}(\mathbf{I}_{n_{\mathcal{A}}} - \mathbf{H}_\lambda)\mathbf{I}_{n_{\mathcal{A}}})^{-2}(\mathbf{I}_{n_{\mathcal{A}}} - \mathbf{H}_\lambda)\mathbf{y} \tag{2.21}$$

$$= \frac{1}{n_{\mathcal{A}}} \mathbf{b}^T(\mathbf{I}_{n_{\mathcal{A}}} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})(\text{tr}(\mathbf{I}_{n_{\mathcal{A}}} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})\mathbf{I}_{n_{\mathcal{A}}})^{-2}(\mathbf{I}_{n_{\mathcal{A}}} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})\mathbf{b}. \tag{2.22}$$

Dans la forme 2.22, toutes les matrices dépendantes de λ sont diagonales et donc très faciles à manipuler. Pour peu que nous disposions de l'ACP de la matrice $\mathbf{ZZ}^T$, la GCV se calcule donc immédiatement.

2.6 Les modèles mixtes

Les modèles à effets aléatoires sont une approche mathématique très proche de la régression ridge. Dans cette section nous allons présenter ces modèles et leur lien avec la régression ridge.

Les modèles à effets aléatoires proposent une approche différente de la modélisation linéaire : contrairement à la régression "classique" (où les coefficients sont pensés comme fixes) les effets sont ici considérés comme des variables aléatoires. Bien que nous ayons (beaucoup) utilisé les modèles à effets aléatoires dans cette thèse, nous avons centré notre travail autour de la régression ridge. Nous ne nous attarderons donc pas sur la théorie des modèles à effets aléatoires et nous nous contenterons d'énoncer les grandes lignes. L'écriture de cette section se base sur les travaux de Dandine-Roulland [2014] et Perdry [2017] qui ont fait un incroyable travail d'explication de ces modèles.

Modèles à effets aléatoires et modèles mixtes

Considérons un modèle linéaire de la forme

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} \tag{2.23}$$

$$= \mathbf{f} + \mathbf{g} + \mathbf{e} \tag{2.24}$$

avec $\beta \in \mathbb{R}^r$, $u \in \mathbb{R}^p$ et $\mathbf{e} \in \mathbb{R}^n$. Dans ce modèle le terme $\mathbf{e} \sim \mathcal{N}(0_n, \sigma^2\mathbf{I}_n)$ joue le rôle d'un vecteur de bruit. Nous remarquons que nos données sont scindées en deux matrices $\mathbf{X}$ et $\mathbf{Z}$. Le vecteur β sera considéré comme un terme fixe contenant des paramètres à estimer. Nous appellerons ces paramètres *effets fixes*. Le vecteur u sera lui considéré comme un vecteur aléatoire $u \sim \mathcal{N}(0_p, \tau\mathbf{I}_p)$. Le paramètre que nous chercherons à estimer est τ et non le vecteur d'effets u. Une approche rigoureusement équivalente est de fixer l'aléatoire sur le terme génétique $\mathbf{g}$ directement $\mathbf{g} \sim \mathcal{N}(\mathbf{0}_n, \tau\mathbf{Z}\mathbf{Z}^T)$.

Dans les deux cas la réponse suit donc une loi normale multivariée,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n). \tag{2.25}$$

Pour estimer les composantes de variances τ et σ^2, nous utiliserons la maximisation de la vraisemblance. En posant $\mathbf{\Sigma} = \tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n$ nous pouvons écrire la log-vraisemblance de 2.23

$$l(\beta, \tau, \sigma^2) = -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\beta) + \text{cst.} \tag{2.26}$$

La matrice $\mathbf{Z}\mathbf{Z}^T$ est parfois appelée la matrice d'apparentement et mesure la ressemblance entre les individus. En dérivant 2.26 selon β, nous obtenons facilement

$$\hat{\beta} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}. \tag{2.27}$$

En remarquant que

$$\mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{\Sigma}\mathbf{P}\mathbf{y}$$

avec $\mathbf{P} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^T\mathbf{X}^T\mathbf{\Sigma}^{-1}$, nous pouvons écrire la log-vraisemblance

profilée du modèle

$$l(\hat{\beta}, \tau, \sigma^2) = -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y} + \text{cst.} \tag{2.28}$$

Un modèle restreint

Nous remarquons que dans la formule 2.26 nous avons besoin de connaître β pour estimer les composantes de variance, mais la formule 2.27 montre également que $\hat{\beta}$ est une fonction des composantes de variance. Pour passer outre ce problème, une solution est d'utiliser une vraisemblance restreinte. Soit $\mathbf{C} \in \mathcal{M}_{n-r,n}$ une matrice dite de contraste qui vérifie $\mathbf{CX} = \mathbf{0}_{n,r}$ et $\mathbf{CC}^T = \mathbf{I}_{n-r}$. En multipliant 2.23 à droite par $\mathbf{C}$ nous obtenons

$$\boldsymbol{\gamma} = \mathbf{Cy} = \mathbf{CX}\beta + \mathbf{CZ}u + \mathbf{Ce} \tag{2.29}$$

$$\rightarrow \boldsymbol{\gamma} = \mathbf{CZ}u + \mathbf{Ce} \sim \mathcal{N}(0_{n-r}, \mathbf{\Omega}) \tag{2.30}$$

avec $\mathbf{\Omega} = \tau\mathbf{CZZ}^T\mathbf{C}^T + \sigma^2\mathbf{I}_{n-r} = \mathbf{C\Sigma C}^T$. Dans ce nouveau modèle il n'y a pas d'effets fixes et les estimations des composantes de variances par maximum de vraisemblance ne seront pas biaisées. La log-vraisemblance de ce modèle est

$$l^{restreinte}(\tau, \sigma^2) = -\frac{1}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{\Omega}\boldsymbol{\gamma} \tag{2.31}$$

En utilisant les deux résultats suivants (que nous admettrons, voir les annexes de Dandine-Roulland [2014] pour une démonstration) des propriétés des matrices de contraste

$$\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{C} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{\Omega}^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}$$

$$\log|\mathbf{\Omega}| = \log|\mathbf{V}| + \log\left|\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\right| + \text{cst}$$

nous obtenons

$$l^{restreinte}(\tau, \sigma^2) = -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\log\left|\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\right| - \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y} + \text{ constante.} \tag{2.32}$$

Estimations des composantes de la variance et des effets génétiques

Il n'existe pas de formule directe des estimateurs du maximum de vraisemblance pour 2.32. L'estimation de τ et σ^2 se fait donc par des méthodes itératives : chaque itération donne une estimation de nos composantes de variance et l'algorithme s'arrête quand les résultats se stabilisent. L'algorithme Espérance-Maximisation (EM) est une solution classiquement utilisée qui converge toujours mais cette convergence peut être très lente. Une méthode souvent utilisée en pratique est l'algorithme *Average Information REscrited Maximum Likelihood* (AI-REML). En notant $\theta = (\tau, \sigma^2)$, le vecteur des composantes estimées à l'itération r est

$$\theta_{r+1} = \theta_r + \frac{1}{2}\left(\mathbf{H}^{-1}(\theta) + \mathbf{I}(\theta)\right)\nabla(\theta)$$

avec $\nabla(\theta)$ le vecteur de gradient de la log-vraisemblance, $\mathbf{H}^{-1}(\theta)$ l'inverse de la matrice Hessienne et $\mathbf{I}(\theta)$ la matrice d'information de Fisher.

Calcul de l'estimateur des effets génétiques

Il est possible de déduire un estimateur du vecteur d'effets génétiques. Rappelons que nous supposons $u \sim \mathcal{N}(0_p, \tau\mathbf{I}_p)$ et $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)$ et remarquons que

$$\begin{bmatrix}\mathbf{y}\\ u\end{bmatrix} = \begin{bmatrix}\mathbf{X}\beta\\ 0_p\end{bmatrix} + \begin{bmatrix}\mathbf{Z} & \mathbf{I}_n\\ \mathbf{I}_p & \mathbf{0}_{p,n}\end{bmatrix},$$

nous pouvons alors écrire

$$\begin{bmatrix}\mathbf{y}\\ u\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{X}\beta\\ 0_p\end{bmatrix}, \begin{bmatrix}\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n & \tau\mathbf{Z}\\ \tau\mathbf{Z}^T & \tau\mathbf{I}_p\end{bmatrix}\right).$$

La distribution de u conditionnellement à $\mathbf{y}$ suit alors une loi normale de paramètre

$$u|\mathbf{y} \sim \mathcal{N}\left(0_p + \tau\mathbf{Z}^T(\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X}\beta), \tau\mathbf{I}_p - \tau\mathbf{Z}^T(\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)^{-1}\tau\mathbf{Z}\right).$$

Le *Best Linear Unbiased Predictor* (abrégé en BLUP) est alors défini comme

$$\hat{u}_{BLUP} = \mathbb{E}[u|\mathbf{y}] = \tau\mathbf{Z}^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\beta). \tag{2.33}$$

Notons que dans 2.33, nous avons utilisé les vraies composantes de variance. Pour estimer ce BLUP nous devrons remplacer les composantes de la variances et le vecteur d'effets fixes par leurs estimations : nous parlerons alors d'*empirical BLUP* (*eBLUP*)

$$\hat{u}_{eBLUP} = \hat{\tau}\mathbf{Z}^T\hat{\mathbf{\Sigma}}(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{2.34}$$

$$= \hat{\tau}\mathbf{Z}^T\hat{\mathbf{P}}\mathbf{y}. \tag{2.35}$$

BLUP et matrice d'apparentement

Jusqu'à présent nous nous sommes surtout intéressés à placer l'aléa sur u et nous avons juste mentionné la possibilité de fixer l'aléa sur $\mathbf{g}$ en disant que les approches étaient équivalentes. Un des intérêts du modèle aléatoire est que les calculs de vraisemblance ne sont pas fonction de la matrice de génotype mais de la matrice d'apparentement $\mathbf{ZZ}^T$. En fixant l'aléa sur $\mathbf{g}$ il est possible d'utiliser n'importe quelle matrice d'apparentement $\mathbf{A}$: nous aurons alors $\mathbf{g} \sim \mathcal{N}(\mathbf{0}_n, \sigma_A^2\mathbf{A})$ et la matrice de covariance de $\mathbf{y}$ devient alors $\mathbf{\Sigma} = \sigma_A^2\mathbf{A} + \sigma^2\mathbf{I}_n$.

Bien sûr sous cette approche il n'est pas possible d'estimer un vecteur d'effets génétiques $\hat{u}$. Nous pourrons quand même estimer le vecteur de composante génétique par

$$\hat{\mathbf{g}} = \sigma_A^2\mathbf{APy}. \tag{2.36}$$

2.6.1 Lien entre modèle à effets aléatoires et régression ridge

La régression ridge et les modèles à effets aléatoires sont deux concepts mathématiques très liés. Pour le voir, intéressons nous à la maximisation de la distribution à posteriori dans le modèle

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e} \tag{2.37}$$

avec $u \sim \mathcal{N}(0_p, \tau\mathbf{I}_p)$, $\mathbf{e} \sim \mathcal{N}(0_n, \sigma^2\mathbf{I}_n)$ et en supposant la matrice de génotypes comme fixée. Notre objectif est de maximiser le *Maximum A Posteriori*

$$p(u|\mathbf{y}) = \frac{p(\mathbf{y}|u)p(u)}{p(\mathbf{y})}$$

$$\rightarrow \log\, p(u|\mathbf{y}) = \log\, p(\mathbf{y}|u) + \log\, p(u) - \log\, p(\mathbf{y}).$$

avec $p(u|\mathbf{y})$ la loi de u conditionné par $\mathbf{y}$. En utilisant le fait que $u \sim \mathcal{N}(0_p, \tau\mathbf{I}_p)$, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}_n, \tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)$, $\mathbf{y}|u \sim \mathcal{N}(\mathbf{Z}u, \sigma^2\mathbf{I}_n)$ et en rappelant que la log-vraisemblance d'une distribution normale multivariée de paramètres μ et $\mathbf{\Sigma}$ vaut

$$\log\, p(\mathbf{x}|\mu, \mathbf{\Sigma}) = -\frac{n}{2}\log\, 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - (\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu),$$

nous pouvons alors écrire

$$\begin{aligned}\log\, p(u|\mathbf{y}) = &-\frac{n}{2}\log\, 2\pi - \frac{1}{2}\log\left|\sigma^2\mathbf{I}_n\right| - \frac{1}{2}(\mathbf{y}-\mathbf{Z}u)^T(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{Z}u)\\ &-\frac{p}{2}\log\, 2\pi - \frac{1}{2}\log|\tau\mathbf{I}_p| - \frac{1}{2}(u-0_p)^T(\tau\mathbf{I}_p)^{-1}(u-0_p)\\ &+\frac{n}{2}\log\, 2\pi + \frac{1}{2}\log\left|\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n\right| - \frac{1}{2}(\mathbf{y}-\mathbf{0}_n)^T(\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{0}_n).\end{aligned}$$

En isolant les termes dépendants de u nous obtenons

$$\log\, p(u|\mathbf{y}) = -\frac{1}{2\sigma^2}\left(\|\mathbf{y}-\mathbf{Z}u\|_2^2 + \frac{\sigma^2}{\tau}\|u\|_2^2\right) + K_{\perp u}$$

avec $K_{\perp u}$ un terme indépendant de u. Au vu de l'équation précédente, nous avons l'équivalence

$$\arg\max_u p\,(u|\mathbf{y}) = \arg\min_u \|\mathbf{y}-\mathbf{Z}u\|_2^2 + \lambda\,\|u\|_2^2 \text{ avec } \lambda = \frac{\sigma^2}{\tau}. \tag{2.38}$$

Il y a donc équivalence entre chercher les composantes de variance optimales et le paramètre de pénalisation optimal d'une régression ridge. Le paramètre de pénalisation est défini comme le ratio des composantes de variance. Pour bien comprendre cette équivalence, nous allons étudier le choix des composantes de variance et du paramètre de pénalisation dans les deux cas extrêmes.

Prenons le cas extrême du modèle intégralement bruité i.e $u = 0_p$. D'après la formule 2.10 nous aurons donc $\forall\lambda$ biais$^2 = 0$ i.e le biais est constant. D'après le dilemme biais-variance, nous allons choisir λ qui minimise la variance et au vu de 2.9 il s'agit de $\lambda = +\infty$. Du point de vue du modèle mixte, $u = 0_p \to \tau = 0$. En utilisant le lien 2.38, nous retrouvons bien $\lambda = +\infty$.

A l'inverse, prenons le cas du modèle sans bruit. Ici $\mathbf{e} = 0 \to \sigma^2 = 0$. D'après 2.9, nous aurons donc la variance constante et nulle pour tout λ. Le dilemme biais variance choisira donc le λ qui minimise le biais i.e $\lambda = 0$. Dans le modèle à effets aléatoires nous sommes dans le cas où $\sigma^2 = 0$. Toujours en utilisant 2.38 nous retrouvons $\lambda = 0$.

Au vu de cette équivalence on se doute que les concepts de BLUP et d'estimateur ridge vont être très proches, ce qui se montre très facilement :

$$\begin{aligned}\hat{u}_R &= \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)\mathbf{Z}^T\mathbf{y} = \mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I}_n\right)^{-1}\mathbf{y} \\ &= \mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T + \frac{\sigma^2}{\tau}\mathbf{I}_n\right)^{-1}\mathbf{y} = \tau\mathbf{Z}^T\left(\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n\right)^{-1}\mathbf{y} \\ &= \hat{u}_{BLUP}.\end{aligned}$$

2.7 Intégration de variables non-pénalisées dans la régression ridge

Jusqu'à présent nous avons supposé que l'on souhaitait pénaliser toutes les variables de notre modèle mais ce ne sera pas toujours le cas. Par exemple nous souhaitons quasiment systématiquement ne pas pénaliser l'intercept de notre modèle.

Introduisons dans le modèle $\mathbf{X} \in \mathcal{M}_{n,r}(\mathbb{R})$ la matrice des variables non pénalisées et $\beta \in \mathbb{R}^r$ le vecteur d'effets de ces variables. Le modèle devient alors

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} \tag{2.39}$$

et le critère ridge associé pour une estimation jointe de β et u sera

$$\underset{\beta\in\mathbb{R}^r,\ u\in\mathbb{R}^p}{\arg\min}\ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}u\|_2^2 + \lambda\|u\|_2^2. \tag{2.40}$$

En dérivant 2.40 selon β et u et en cherchant l'annulation des dérivées, nous obtenons

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{Z}u) \\ \hat{u}_R &= \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\beta).\end{aligned}$$

Ces formules ne sont pas utilisables car en pratique nous ne connaissons pas β et u. Si nous sommes intéressés par l'estimation de u uniquement, nous utiliserons des matrices de projection.

Avec le projecteur orthogonal

Le choix de matrice de projection le plus classique est

$$\mathbf{P}_{\mathbf{X}} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2.41}$$

avec $\mathbf{P}_{\mathbf{X}}$ qui vérifie en particulier $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{0}_n$, $\mathbf{P}_{\mathbf{X}}^T = \mathbf{P}_{\mathbf{X}}$ et $\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}}$.

Alors en multipliant 2.39 à gauche par 2.41, nous obtenons le modèle sans covariables non pénalisées

$$\mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{P}_{\mathbf{X}}\mathbf{X}\beta + \mathbf{P}_{\mathbf{X}}\mathbf{Z}u + \mathbf{P}_{\mathbf{X}}\mathbf{e} = \mathbf{P}_{\mathbf{X}}\mathbf{Z}u + \mathbf{P}_{\mathbf{X}}\mathbf{e} \tag{2.42}$$

En particulier dans ce nouveau modèle nous avons

$$\mathbf{P}_{\mathbf{X}}\mathbf{y} \sim \mathcal{N}(\mathbf{P}_{\mathbf{X}}\mathbf{Z}u, \sigma^2\mathbf{P}_{\mathbf{X}}). \tag{2.43}$$

Pour obtenir un estimateur de u on écrit le critère ridge du modèle (2.42)

$$\begin{aligned} &\arg\min_{u\in\mathbb{R}^p} \|\mathbf{P}_{\mathbf{X}}\mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{Z}u\|_2^2 + \lambda \|u\|_2^2 && (2.44)\\ =&\arg\min_{u\in\mathbb{R}^p} (\mathbf{y} - \mathbf{Z}u)^T\mathbf{P}_{\mathbf{X}}(\mathbf{y} - \mathbf{Z}u) + \lambda u^T u && (2.45) \end{aligned}$$

et en dérivant ce critère selon u puis en cherchant à annuler la dérivée nous obtenons

$$\begin{aligned} \hat{u}_R &= \left(\mathbf{Z}^T\mathbf{P}_{\mathbf{X}}\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{P}_{\mathbf{X}}\mathbf{y} && (2.46)\\ &= \mathbf{Z}^T\mathbf{P}_{\mathbf{X}}(\mathbf{P}_{\mathbf{X}}\mathbf{Z}\mathbf{Z}^T\mathbf{P}_{\mathbf{X}} + \lambda\mathbf{I}_n)\mathbf{P}_{\mathbf{X}}\mathbf{y}. && (2.47) \end{aligned}$$

Avec une matrice de contraste

Il est également possible de travailler dans un espace réduit comme dans les modèles mixtes. Notons $\mathbf{C}$ une matrice de contraste vérifiant $\mathbf{C}\mathbf{X} = \mathbf{0}_{n,r}$ et $\mathbf{C}\mathbf{C}^T = \mathbf{I}_{n-r}$, alors comme pour l'approche précédente en multipliant 2.39 à droite par $\mathbf{C}$ nous obtenons un modèle sans variables non-pénalisées

$$\mathbf{Cy} = \mathbf{CX}\beta + \mathbf{CZ}u + \mathbf{Ce} = \mathbf{CZ}u + \mathbf{Ce}. \tag{2.48}$$

Une différence par rapport à l'approche précédente est que

$$\mathbf{Cy} \sim \mathcal{N}(\mathbf{CZ}u, \sigma^2\mathbf{I}_{n-r}) \tag{2.49}$$

i.e la matrice de covariance du phénotype contrasté est une matrice diagonale à coefficients constants. Nous écrivons le critère ridge du modèle (2.48)

$$\arg\min_{u\in\mathbb{R}^p} \|\mathbf{Cy} - \mathbf{CZ}u\|_2^2 + \lambda \|u\|_2^2 \tag{2.50}$$

$$= \arg\min_{u\in\mathbb{R}^p} (\mathbf{y} - \mathbf{Z}u)^T\mathbf{C}^T\mathbf{C}(\mathbf{y} - \mathbf{Z}u) + \lambda u^T u \tag{2.51}$$

et nous obtenons après dérivation

$$\hat{u}_R = \left(\mathbf{Z}^T\mathbf{C}^T\mathbf{CZ} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{C}^T\mathbf{Cy} \tag{2.52}$$

$$= \mathbf{Z}^T\mathbf{C}^T(\mathbf{CZZ}^T\mathbf{C}^T + \lambda\mathbf{I}_{n-r})\mathbf{Cy}. \tag{2.53}$$

Choix d'une matrice de contraste

La question qui se pose est comment trouver $\mathbf{C}$? Une approche est d'utiliser la décomposition QR. La décomposition QR de $\mathbf{A} \in \mathbb{M}_{n,r}(\mathbb{R})$ s'écrit $\mathbf{A} = \mathbf{QR}$ avec $\mathbf{Q} \in \mathcal{O}(n)$ et $\mathbf{R}^T \in \mathcal{M}_{r,n}(\mathbb{R}) = [\mathbf{R}_1^T, \mathbf{0}_{r,n-r}]$ où $\mathbf{R}_1 \in \mathcal{M}_{r,r}(\mathbb{R})$ est une matrice triangulaire supérieure. En posant $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$ avec $\mathbf{Q}_1 \in \mathcal{M}_{n,r}(\mathbb{R})$, $\mathbf{Q}_2 \in \mathcal{M}_{n,n-r}(\mathbb{R})$ et en remarquant que

$$\mathbf{Q}^T\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1^T\\ \mathbf{Q}_2^T\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix} = \begin{pmatrix}\mathbf{Q}_1^T\mathbf{Q}_1 & \mathbf{Q}_1^T\mathbf{Q}_2\\ \mathbf{Q}_2^T\mathbf{Q}_1 & \mathbf{Q}_2^T\mathbf{Q}_2\end{pmatrix} = \begin{pmatrix}\mathbf{I}_r & \mathbf{O}_{r,n-r}\\ \mathbf{O}_{n-r,r} & \mathbf{I}_{n-r}\end{pmatrix},$$

alors nous montrons que l'on a

$$\mathbf{Q}_2^T\mathbf{A} = \mathbf{Q}_2^T\begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}\mathbf{R} = \begin{pmatrix}\mathbf{0}_{n-r,r} & \mathbf{I}_{n-r}\end{pmatrix}\begin{pmatrix}\mathbf{R}_1 \\ \mathbf{0}_{n-r,n}\end{pmatrix} = \mathbf{0}_{n-r,r}.$$

Comme $\mathbf{Q}_2^T\mathbf{A} = \mathbf{0}_{n-r,r}$ et $\mathbf{Q}_2^T\mathbf{Q}_2 = \mathbf{I}_{n-r}$, alors c'est une matrice de contraste. La décomposition QR d'une matrice étant relativement peu coûteuse à calculer, cette méthode est bien pratique.

Estimation disjointe

Dans la pratique plutôt que d'utiliser les approches décrites au dessus, il sera parfois tentant d'effectuer une estimation disjointe pour les variables pénalisées et non-pénalisées (particulièrement dans les cas où nous nous attendons à ce que les variables pénalisées soient indépendantes des non-pénalisées). Un exemple d'un tel cas pourrait être un modèle avec un intercept très élevé. Nous commençons par estimer les effets fixes avec $\mathbf{y}$ comme réponse

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

Ensuite, nous calculons l'estimateur ridge avec comme réponse la réponse $\mathbf{y}$ moins l'estimation du terme des effets non-pénalisés

$$\hat{u}_R = \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\beta}).$$

2.8 Extension du lien entre régression ridge et modèle à effets aléatoires en présence d'effets fixes

Il est possible de généraliser le lien établi dans la section 2.6.1 au cas avec effets fixes. Prenons le modèle défini en (2.39)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e}$$

avec $u \sim \mathcal{N}(0_p, \tau\mathbf{I}_p)$ et $\mathbf{e} \sim \mathcal{N}(0_n, \sigma^2\mathbf{I}_n)$. En remarquant que $\mathbf{y}|u \sim \mathcal{N}(\mathbf{X}\beta + \mathbf{Z}u, \sigma^2\mathbf{I}_n)$, en supposant encore une fois $\mathbf{Z}$ fixée et en utilisant les mêmes calculs que plus haut

nous pouvons montrer l'équivalence

$$\arg\max_u p\left(u|\mathbf{y}\right) = \arg\min_u \left\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}u\right\|_2^2 + \lambda \left\|u\right\|_2^2 \text{ avec } \lambda = \frac{\sigma^2}{\tau}. \tag{2.54}$$

Notons également que le lien se prolonge également au cas des modèles contrastés. Prenons la version contrastée de (2.39)

$$\mathbf{C}\mathbf{y} = \mathbf{C}\mathbf{Z}u + \mathbf{C}\mathbf{e}. \tag{2.55}$$

avec $\mathbf{C}$ une matrice de contraste de $\mathbf{X}$. En remarquant que $\mathbf{C}\mathbf{e} \sim \mathcal{N}(\mathbf{0}_{n-r}, \sigma^2\mathbf{I}_{n-r})$ et que $\mathbf{C}\mathbf{y} \sim \mathcal{N}(\mathbf{C}\mathbf{Z}u, \sigma^2\mathbf{I}_{n-r})$, nous avons alors l'équivalence

$$\arg\max_u p\left(u|\mathbf{C}\mathbf{y}\right) = \arg\min_u \left\|\mathbf{C}\mathbf{y} - \mathbf{C}\mathbf{Z}u\right\|_2^2 + \lambda \left\|u\right\|_2^2 \text{ avec } \lambda = \frac{\sigma^2}{\tau}. \tag{2.56}$$

Notons en particulier que dans le modèle initial (2.39) ou dans le modèle contrasté le paramètre de pénalisation optimal est le même, ce qui remontre l'intérêt de travailler dans un modèle contrasté.

Chapitre 3

Validation croisée généralisée en grande dimension

Utiliser la GCV en grande dimension n'a pas été immédiat. En effet cette dernière avait tendance à très fortement sous-estimer le paramètre de pénalisation optimal car il était systématiquement nul. Dans cette section nous décrirons les problèmes que nous avons rencontrés et expliquerons comment nous avons réussi à les résoudre.

Nous travaillerons avec le modèle linéaire classique.

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e} \tag{3.1}$$

Avec $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$, $u \in \mathbb{R}^p$ et $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. Nous supposerons que toutes les variables de la matrice de données sont pénalisées. Ainsi l'erreur de GCV pour choisir le paramètre de pénalisation λ de la régression ridge s'écrit

$$\begin{aligned} \mathrm{err}^{GCV}(\lambda) &= \frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{H}_\lambda)(\mathrm{tr}(\mathbf{I}_n - \mathbf{H}_\lambda)\mathbf{I}_n)^{-2}(\mathbf{I}_n - \mathbf{H}_\lambda)\mathbf{y} \\ &= \frac{1}{n}\mathbf{b}^T(\mathbf{I}_n - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})(\mathrm{tr}(\mathbf{I}_n - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})\mathbf{I}_n)^{-2}(\mathbf{I}_n - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}})\mathbf{b}. \end{aligned}$$

avec

$$\boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}} = \mathbf{D}\mathbf{D}^T\left(\mathbf{D}\mathbf{D}^T + \lambda\mathbf{I}_n\right)^{-1} \text{ et} \tag{3.2}$$

$$\mathbf{b} = \mathbf{U}^T\mathbf{y}. \tag{3.3}$$

Nous supposerons également que le centrage et la réduction de $\mathbf{Z}$ seront réalisés de manière empirique. Soit $\mathbf{G} \in \mathcal{M}_{n,p}(\mathbb{N})$ la matrice de données brutes. Nous calculerons pour chaque variant j la moyenne empirique

$$\hat{\mu}_j = \frac{1}{n}\mathbb{1}_n^T \mathbf{G}_j = \frac{1}{n}\sum_{i=1}^{n} G_{i,j}$$

et l'écart-type empirique

$$\hat{\sigma}_j = \frac{1}{n-1}(\mathbf{G}_j - \hat{\mu}_j \mathbb{1}_n)^T(\mathbf{G}_j - \hat{\mu}_j \mathbb{1}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(G_{i,j} - \hat{\mu}_j)^2.$$

Alors la matrice de données standardisées $\mathbf{Z}$ est définie comme

$$\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_p] \text{ avec } \mathbf{z}_j = (\mathbf{G}_j - \hat{\mu}_j \mathbb{1}_n)/\hat{\sigma}_j.$$

3.1 Illustration des problèmes de la GCV en grande dimension et sous centrage empirique

Nous allons illustrer les problèmes de la GCV en grande dimension et avec un mauvais choix de centrage en utilisant des simulations qui seront résumées dans le graphe 3.1. Nous avons simulé $n = 100$ individus selon un modèle linéaire pour un scénario "petite dimension" avec $p = 20$ (panneaux du haut) et un scénario "grande dimension" avec $p = 2000$ (panneaux du bas). Nous comparerons le centrage empirique de la matrice des génotypes (panneaux de gauche) et un centrage avec des valeurs externes qui seront ici les valeurs que nous avons utilisées pour la simulation des génotypes (panneaux de droite). Dans tous ces scénarios les phénotypes seront centrés empiriquement. Notre objectif est de chercher le paramètre de pénalisation optimal pour la régression ridge. Pour cela nous avons calculé la GCV et également estimé la précision du modèle par $\frac{1}{p}(u - \hat{u}_R)^T(u - \hat{u}_R)$ avec u le vecteur d'effet simulé et $\hat{u}_R$ l'estimateur de la régression ridge. Dans chacun des panneaux du graphe 3.1 nous avons affiché l'erreur de GCV et la précision de l'estimateur en fonction des degrés de liberté effectifs définis en (1.15). Toutes les courbes de GCV ont été multipliées par une constante pour être à la même échelle que la précision mais ce n'est pas un problème car nous nous intéressons uniquement aux minimums des courbes et non pas à leurs valeurs.

Nous constatons que dans le cas $n > p$ la GCV et la précision choisissent la même complexité optimale qu'importe le centrage. Cela montre qu'en petite dimension utiliser

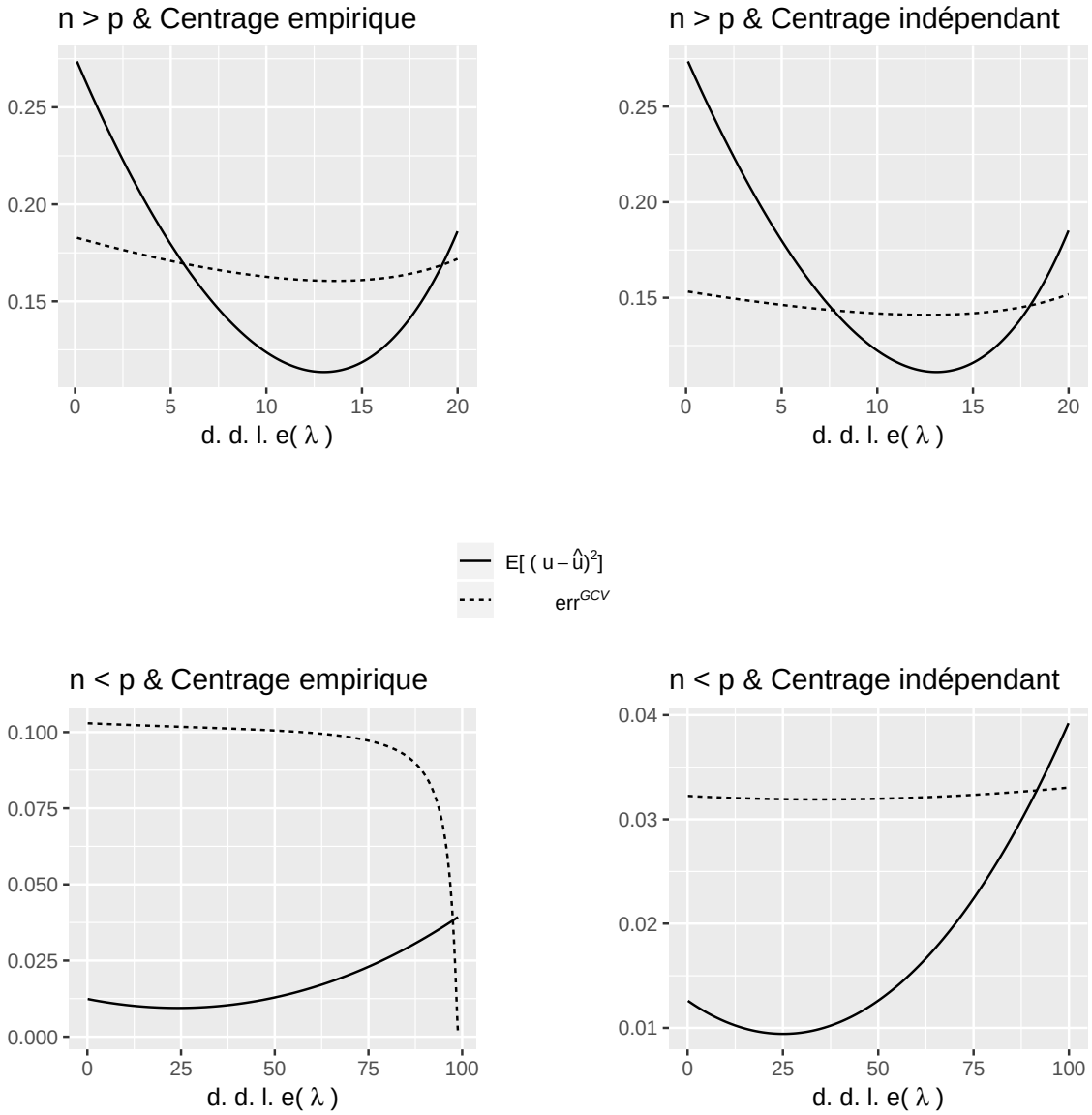


FIGURE 3.1 – Une illustration du biais de la GCV en grande dimension et avec le centrage empirique. Chaque panel qui correspond à un scénario de dimension et de centrage représente la précision de l'estimateur (courbe pleine) et l'erreur de la GCV (courbe en pointillés) selon les degrés de liberté effectifs comme mesure de la complexité. Les deux études sont de taille $n = 100$ avec $p = \{20, 2000\}$. Le centrage indépendant a été réalisé avec des valeurs issues des simulations.

un centrage empirique n'a pas un effet net sur la qualité de la GCV.

En revanche dans le cas $n < p$ le choix de standardisation se révèle avoir une grande importance. Dans le panel avec standardisation par données externes, la GCV et la précision choisissent une complexité optimale assez proche. Avec une standardisation empirique la GCV a un comportement étrange et choisit comme complexité optimale le cas où $\lambda = 0$ (autrement dit le modèle le plus complexe), et a un comportement différent de la précision.

Ce comportement est inévitable. Dans la section suivante, nous allons montrer que cela est dû à la conjonction de plusieurs facteurs : le contexte de la grande dimension, de standardisation empirique de la matrice des génotypes et enfin l'estimation disjointe de l'intercept par la moyenne empirique et des effets génétiques.

3.2 Démonstrations des problèmes

Commençons par remarquer que le spectre de la matrice $\mathbf{K} = \mathbf{Z}\mathbf{Z}^T$ est différent selon que l'on soit en petite ou en grande dimension. Nous rappelons l'expression de la SVD de $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T \rightarrow \mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{D}^T\mathbf{U}^T$, nous avons alors

— En petite dimension, $\mathbf{D}\mathbf{D}^T = \left(\begin{array}{ccc|c} d_1^2 & & & \\ & \ddots & & \mathbf{0}_{p,n-p} \\ & & d_p^2 & \\ \hline \multicolumn{3}{c|}{\mathbf{0}_{n-p,p}} & \mathbf{0}_{n-p} \end{array}\right)$ et $\text{sp}(\mathbf{K}) = \{d_1^2, ..., d_p^2, 0\}$.

— En grande dimension, $\mathbf{D}\mathbf{D}^T = \begin{pmatrix} d_1^2 & & \\ & \ddots & \\ & & d_n^2 \end{pmatrix}$ et $\text{sp}(\mathbf{K}) = \{d_1^2, ..., d_n^2\}$.

Remarquons désormais l'influence d'un centrage empirique des génotypes :

$$\forall j \in [\![1, p]\!],\ \sum_{i=1}^{n} z_{i,j} = 0 \rightarrow z_i = -\sum_{j \neq i} z_j.$$

avec $z_i \in \mathbb{R}^p$ un vecteur colonne représentant la i-ème ligne de $\mathbf{Z}$.

Avec ce choix de centrage chaque ligne de $\mathbf{Z}$ est une combinaison linéaire de toutes les autres. Une manière de le voir est que l'un de nos individus n'est plus informatif, puis qu'il est parfaitement décrit par tous les autres. Cela a une influence sur l'ACP

de $\mathbf{K}$. Rappelons que l'ACP translate notre jeu de données pour créer des variables les plus informatives possibles. Comme le centrage empirique "sacrifie" un individu, une conséquence est que l'une des valeurs propres de $\mathbf{K}$ est automatiquement mise à 0 : le centrage empirique fait "perdre une dimension" à l'ACP.

Cette perte de dimension a une influence différente entre la petite et la grande dimension : en petite dimension, l'ACP a déjà au maximum p directions explicatives et donc $n - p$ directions non explicatives avec une valeur propre associé nulle. En conséquence la perte d'une dimension par le centrage n'est pas la cause de l'introduction de la valeur propre nulle dans le spectre de $\mathbf{K}$.

Au contraire en grande dimension les n directions de l'ACP sont censées être explicatives : le fait que l'on force $d_n^2 = 0$ va significativement changer le spectre de $\mathbf{K}$.

Nous allons maintenant démontrer de deux manières que la GCV choisit $\lambda = 0$ dans le contexte de grande dimension et de centrage empirique. La première démonstration est spécifique à la GCV et la deuxième est plus générale pour la Leave-One-Out avec standardisation unique pour tout les individus définie en (2.18).

Démonstration 1

Commençons par montrer qu'en cas de centrage empirique de $\mathbf{Z}$ la valeur propre nulle est associée aux vecteurs propres à coefficients identiques : soit $\mathbf{w} = \alpha \mathbb{1}_{\boldsymbol{n}}$ avec $\alpha \in \mathbb{R}$. Puisque $\mathbf{Z}$ est centré empirique, nous avons $\mathbf{Z}^T\mathbf{w} = 0_p$ et donc $\mathbf{Z}\mathbf{Z}^T\mathbf{w} = \mathbf{0}_{\boldsymbol{n}} = 0\mathbf{w}$ i.e. tout vecteur à coefficients constants est un vecteur propre de $\mathbf{K}$ associé à 0. En prenant $\alpha = \frac{1}{\sqrt{n}}$ ou $\alpha = \frac{-1}{\sqrt{n}}$, alors $\mathbf{w}$ est un vecteur propre et unitaire.

Nous remarquons que en grande dimension et avec un centrage empirique nous avons

$$\mathbf{I}_n - \boldsymbol{\mathcal{D}}_\lambda \xrightarrow[\lambda \to 0]{d_n^2 = 0} \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & 1 \end{pmatrix}$$

avec $\boldsymbol{\mathcal{D}}_\lambda$ définie comme en (3.2) et donc

$$\left[\frac{1}{n}\mathrm{tr}(\mathbf{I}_n - \boldsymbol{\mathcal{D}}_\lambda)\mathbf{I}_n\right]^{-2} \xrightarrow[\lambda\to 0]{d_n^2=0} n^2\mathbf{I}_n.$$

En utilisant ces deux résultats, nous avons donc

$$\mathrm{err}^{GCV}(\lambda) \xrightarrow[\lambda\to 0]{d_n^2=0} \frac{1}{n} \times n^2\mathbf{b}^T \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & 1 \end{pmatrix} \mathbf{b} = nb_n^2.$$

Nous rappelons que $\mathbf{b} = \mathbf{U}^T\mathbf{y}$ i.e. la n-ième colonne de $\mathbf{U}$ correspond au vecteur propre associé à la valeur propre nulle. Comme $\mathbf{K}$ admet une valeur propre nulle et que nous venons de voir qu'en cas de standardisation empirique la valeur propre nulle a l'un de ces vecteurs propres associés qui est à coefficients constants, alors la dernière colonne de $\mathbf{U}$ est à coefficients constants. Nous avons donc

$$b_n^2 = \frac{1}{n}(\mathbb{1}_{\boldsymbol{n}}{}^T\mathbf{y})^2$$

et donc

$$\mathrm{err}^{GCV}(\lambda) \xrightarrow[\lambda\to 0]{d_n^2=0} (\mathbb{1}_{\boldsymbol{n}}{}^T\mathbf{y})^2.$$

Si nous décidons de travailler sur un phénotype centré empiriquement, alors $\mathbb{1}_{\boldsymbol{n}}{}^T\mathbf{y} = 0$ et donc

$$\mathrm{err}^{GCV}(\lambda) \xrightarrow[\lambda\to 0]{d_n^2=0} 0.$$

Travailler sur un phénotype centré empiriquement revient à travailler dans un modèle avec un terme d'intercept puis à estimer de façon disjointe l'intercept et les effets aléatoires. C'est une pratique très courante en petite dimension sans que cela ne pose de problème. Nous avons donc montré que la GCV n'est pas fonctionnelle dans ce contexte de grande dimension avec centrage empirique des données.

Démonstration 2

Dans cette démonstration nous allons remarquer qu'un centrage empirique de $\mathbf{Z}$ (et $\mathbf{y}$) implique que chaque ligne de $\mathbf{Z}$ (et $\mathbf{y}$) devient une combinaison linéaire des autres lignes. Une conséquence pour la LOO est que l'individu de validation devient une combinaison linéaire de tous les individus de l'ensemble d'apprentissage. Nous brisons donc l'indépendance voulue entre les ensembles d'apprentissage et de validation, ce qui explique un biais dans le choix du paramètre de pénalisation.

On rappelle qu'avec un centrage empirique des données nous avons $\forall i \in [\![1, n]\!]$

$$y_i = -\sum_{k \neq i} y_k = -\mathbb{1}_{n-1}^T \mathbf{y}_{-i},$$
$$z_i = -\sum_{k \neq i} z_k = -\mathbb{1}_{n-1}^T \mathbf{Z}_{-i}.$$

Intéressons nous à

$$\hat{\mathbf{y}}_{-i}(i) = z_i^T \hat{u}_R^{-i} = -\sum_{k \neq i} z_k^T \hat{u}_R^{-i} = -\mathbb{1}_n^T \mathbf{Z}_{-i} \mathbf{Z}_{-i}^T \left(\mathbf{Z}_{-i} \mathbf{Z}_{-i}^T + \lambda \mathbf{I}_{n-1} \right)^{-1} \mathbf{y}_{-i}$$

qui est l'estimation pour l'individu i avec l'estimateur ridge construit en excluant i. Nous remarquons que si $\mathbf{Z}_{-i}\mathbf{Z}_{-i}^T$ est inversible (normalement elle l'est en grande dimension car le centrage utilise l'individu i en plus), alors quand $\lambda \to 0$ on a

$$\hat{\mathbf{y}}_{-i}(i) = -\sum_{j \neq i} y_j = y_i.$$

Nous en déduisons que

$$\lambda = 0 \Rightarrow \left(\hat{\mathbf{y}}_{-i}(i) - y_i \right)^2 = 0$$

et donc que

$$\mathrm{err}^{LOO}(\lambda = 0) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\mathbf{y}}_{-i}(i) \right)^2 = 0.$$

Une comparaison avec la petite dimension

A titre de comparaison nous nous proposons de regarder comment se comporte la GCV en petite dimension.

Remarquons que

$$\mathbf{I}_n - \boldsymbol{\mathcal{D}}_\lambda \xrightarrow[\lambda \to 0]{} \begin{pmatrix} \mathbf{0}_p & \mathbf{0}_{p,n-p} \\ \mathbf{0}_{n-p,p} & \mathbf{I}_{n-p} \end{pmatrix}$$

et que donc

$$\mathrm{err}^{GCV}(\lambda) \xrightarrow[\lambda \to 0]{} \left(\frac{1}{n} \times (n-p)\right)^{-2} \mathbf{b}^T \begin{pmatrix} \mathbf{0}_p & \mathbf{0}_{p,n-p} \\ \mathbf{0}_{n-p,p} & \mathbf{I}_{n-p} \end{pmatrix} \mathbf{b} = \left(\frac{n}{n-p}\right)^2 \sum_{k=p+1}^{n} b_k^2.$$

Cette somme n'a aucune raison d'être nulle, même si le terme correspondant au vecteur propre constant est nul. Cela explique pourquoi nous ne retrouvons pas les mêmes biais qu'en grande dimension. Si nous souhaitons nous rassurer sur le fait que la somme soit non nulle, remarquons qu'en notant $\mathbf{b}_r = [\mathbf{U}_{p+1}, ..., \mathbf{U}_n]^T \mathbf{y} = \mathbf{U}_r^T \mathbf{y} \in \mathbb{R}^{n-p}$, alors on a

$$\mathbf{b}_r \sim \mathcal{N}(\mathbf{U}_r^T \mathbf{Z} u, \sigma^2 \mathbf{U}_r^T \mathbf{U}_r) \to \mathbf{b}_r \sim \mathcal{N}(0_{n-p}, \sigma^2 \mathbf{I}_{n-p}).$$

Nous pouvons alors voir cette somme comme proportionnelle à l'estimateur empirique de la variance de $\mathbf{b}_r$, qui est donc non nulle.

3.3 Deux corrections pour la GCV

3.3.1 Une correction avec une matrice de contraste

Le cas général

Puisque l'estimation disjointe de l'intercept pose problème, une solution va être d'utiliser une matrice de contraste pour retirer le terme d'intercept. Nous avons vu dans 2.6.1 que le paramètre de pénalisation était le même dans un modèle avec effets fixes ou dans sa version contrastée. Nous allons donc simplement calculer la GCV sur le modèle contrasté. Prenons le modèle général avec effets fixes (dont l'intercept) : en notant $\mathbf{X} \in \mathcal{M}_{n,r}(\mathbb{R})$ la matrice des covariables non pénalisées et $\beta \in \mathbb{R}^r$ le vecteur d'effet de ces covariables, le modèle s'écrit

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e}. \tag{3.4}$$

Notons $\mathbf{C} \in \mathcal{M}_{n-r,n}(\mathbb{R})$ une matrice de contraste pour les effets fixes, le modèle devient alors

$$\mathbf{y}_C = \mathbf{Z}_C u + \mathbf{e}_C \tag{3.5}$$

avec $\mathbf{y}_C = \mathbf{C}\mathbf{y} \in \mathbb{R}^{n-r}$, $\mathbf{Z}_C = \mathbf{C}\mathbf{Z} \in \mathcal{M}_{n-r,p}(\mathbb{R})$ et $\mathbf{e}_C = \mathbf{C}\mathbf{e} \in \mathbb{R}^{n-r}$. En écrivant $\mathbf{Z}_C = \mathbf{U}_C\mathbf{D}_C\mathbf{V}_C^T$ la SVD de $\mathbf{Z}_C$,

$$\boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}^{\boldsymbol{C}} = \mathbf{D}_C\mathbf{D}_C^T(\mathbf{D}_C\mathbf{D}_C^T + \lambda\mathbf{I}_{n-r})^{-1} \text{ et} \tag{3.6}$$

$$\mathbf{b}_C = \mathbf{U}_C\mathbf{y}_C, \tag{3.7}$$

nous pouvons alors écrire l'erreur de GCV dans le modèle contrasté

$$\mathrm{err}_C^{GCV}(\lambda) = \frac{1}{n-r}\mathbf{b}_C^T(\mathbf{I}_{n-r} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}^{\boldsymbol{C}})(\mathrm{tr}(\mathbf{I}_{n-r} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}^{\boldsymbol{C}})\mathbf{I}_{n-r})^{-2}(\mathbf{I}_{n-r} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\lambda}}^{\boldsymbol{C}})\mathbf{b}_C.$$

Cette formule de GCV contrastée nécessite le calcul d'une matrice de contraste puis le calcul de l'ACP. Dans le cas où les effets fixes se limitent à l'intercept, ce calcul peut être grandement simplifié.

Le cas particulier de l'intercept

Un intercept est par définition un vecteur constant. Par définition de la matrice de contraste, nous chercherons donc une matrice $\mathbf{C} \in \mathcal{M}_{n-1,n}(\mathbb{R})$ vérifiant $\mathbf{C}\mathbb{1}_n = \mathbf{0}_{n-1}$ et $\mathbf{C}\mathbf{C}^T = \mathbf{I}_{n-1}$. Puisque nous avons vu que la dernière colonne de $\mathbf{U}$ était constante, alors en notant $\mathbf{U}_{-n} = [\mathbf{U}_1, ..., \mathbf{U}_{n-1}]$ nous avons $\mathbf{U}_{-n}^T\mathbb{1}_n = \mathbf{0}_{n-1}$ et $\mathbf{U}_{-n}^T\mathbf{U}_{-n} = \mathbf{I}_{n-1}$ par propriété des matrices orthogonales.

L'utilisation de $\mathbf{U}_{-n}^T$ comme matrice de contraste donne à la GCV une forme très rapide à calculer. En effet, nous avons

$$\mathbf{Z}_{\mathbf{U}_{-n}^T} = \mathbf{U}_{-n}^T\mathbf{Z} = \mathbf{U}_{-n}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{D}_{-n}\mathbf{V}^T$$

avec $\mathbf{D}_{-n}$ la matrice $\mathbf{D}$ privée de la ligne n. Il en découle

$$\mathbf{Z}_{\mathbf{U}^T_{-n}}\mathbf{Z}^T_{\mathbf{U}^T_{-n}} = \mathbf{D}_{-n}\mathbf{D}^T_{-n} \in \mathbb{D}_{n-1}(\mathbb{R}) \tag{3.8}$$

$$\rightarrow \mathbf{H}_{\mathbf{U}^T_{-n},\lambda} = \mathbf{Z}_{\mathbf{U}^T_{-n}}\mathbf{Z}^T_{\mathbf{U}^T_{-n}}(\mathbf{Z}_{\mathbf{U}^T_{-n}}\mathbf{Z}^T_{\mathbf{U}^T_{-n}} + \lambda\mathbf{I}_{n-1})^{-1} \in \mathbb{D}_{n-1}(\mathbb{R}). \tag{3.9}$$

Avec ce choix de matrice de contraste, la matrice chapeau $\mathbf{H}_{\mathbf{U}^T_{-n},\lambda}$ associée au modèle est une matrice diagonale. Dans ce modèle constrasté, l'erreur de GCV s'écrit alors

$$\begin{aligned}\mathrm{err}^{GCV}_{\mathbf{U}^T_{-n}}(\lambda) &= \frac{1}{n-1}\mathbf{b}^T_{\mathbf{U}^T_{-n}}(\mathbf{I}_{n-1} - \mathbf{H}_{\mathbf{U}^T_{-n},\lambda})(\mathrm{tr}(\mathbf{I}_{n-1} - \mathbf{H}_{\mathbf{U}^T_{-n},\lambda})\mathbf{I}_{n-1})^{-2}(\mathbf{I}_{n-1} - \mathbf{H}_{\mathbf{U}^T_{-n},\lambda})\mathbf{b}_{\mathbf{U}^T_{-n}} \\ &= \frac{1}{n-1}\frac{\sum_{k=1}^{n-1}\left(\frac{\lambda}{d_k+\lambda}\right)^2 b_k^2}{\left(\frac{1}{n-1}\sum_{k=1}^{n-1}\frac{\lambda}{d_k+\lambda}\right)^2}\end{aligned}$$

Nous retrouvons une formule très proche de la GCV standard (2.19) mais sans le terme associé à la valeur propre nulle dans les sommes.

3.3.2 Une correction brisant les dépendances avec un deuxième ensemble de données

Nous avons montré plus haut que l'origine des problèmes de la GCV vient de la combinaison du cadre de la grande dimension et du centrage empirique, ce qui entraîne des dépendances. Une solution pour supprimer ces dépendances est de réaliser un centrage avec des valeurs indépendantes de l'ensemble d'apprentissage. Nous proposons d'utiliser un ensemble de donnés indépendant (que nous appellerons ensemble de standardisation) pour apprendre les moyennes et écarts-types des variants mais également l'intercept et les autres covariables du modèle. Nous utiliserons ensuite ces estimateurs sur l'ensemble d'apprentissage, sur lequel nous pourrons utiliser la GCV sans faire apparaître de dépendances.

Cette méthode présente deux défauts. Le premier est qu'elle nous oblige à faire une hypothèses d'indépendance entre les variables non-pénalisées et le terme génétiques des variants puisque nous réalisons une estimation en deux temps. Le deuxième défaut est que cette approche est plus gourmande en données puisque nous allons devoir affecter un nombre suffisant d'individus à l'ensemble de standardisation.

Le graphe 3.2 montre un exemple de GCV corrigée pour un exemple en grande

dimension et avec centrage empirique. Nous avons repris les simulations du panel du bas à gauche de la figure 3.1. Nous avons représenté la précision, l'erreur de GCV réparée avec une matrice de contraste et l'erreur de GCV réparée avec un centrage non empirique. Pour la GCV contrastée nous avons utilisé $\mathbf{U}_{-n}^T$ comme matrice de contraste. Pour la correction avec centrage non-empirique nous avons simulé 1000 individus sur lesquels nous avons estimé moyennes et écarts-types empiriques. Nous avons ensuite utilisé ces moyennes et écarts-type pour standardiser l'ensemble d'apprentissage.

Nous voyons que les erreurs de GCV ne semblent pas être monotones décroissantes ni donc toujours choisir $\lambda = 0$. Elles ne sont donc plus biaisées et peuvent donc à priori être considérées pour choisir le paramètre de pénalisation. Si nous les comparons à la précision, les trois courbes semblent avoir un minimum assez proche.

Nous avons exposé et démontré dans ce chapitre les biais que pouvait rencontrer la GCV dans un cadre de grande dimension et avec un centrage empirique des données. Nous avons proposé deux corrections, une basée sur des matrices de contraste et l'autre sur un jeu de données externes pour éviter l'apparition de dépendance. Nous évaluerons de manière plus exhaustive la capacité des GCV corrigées à trouver la complexité optimale du modèle avec des simulations dans le chapitre suivant.

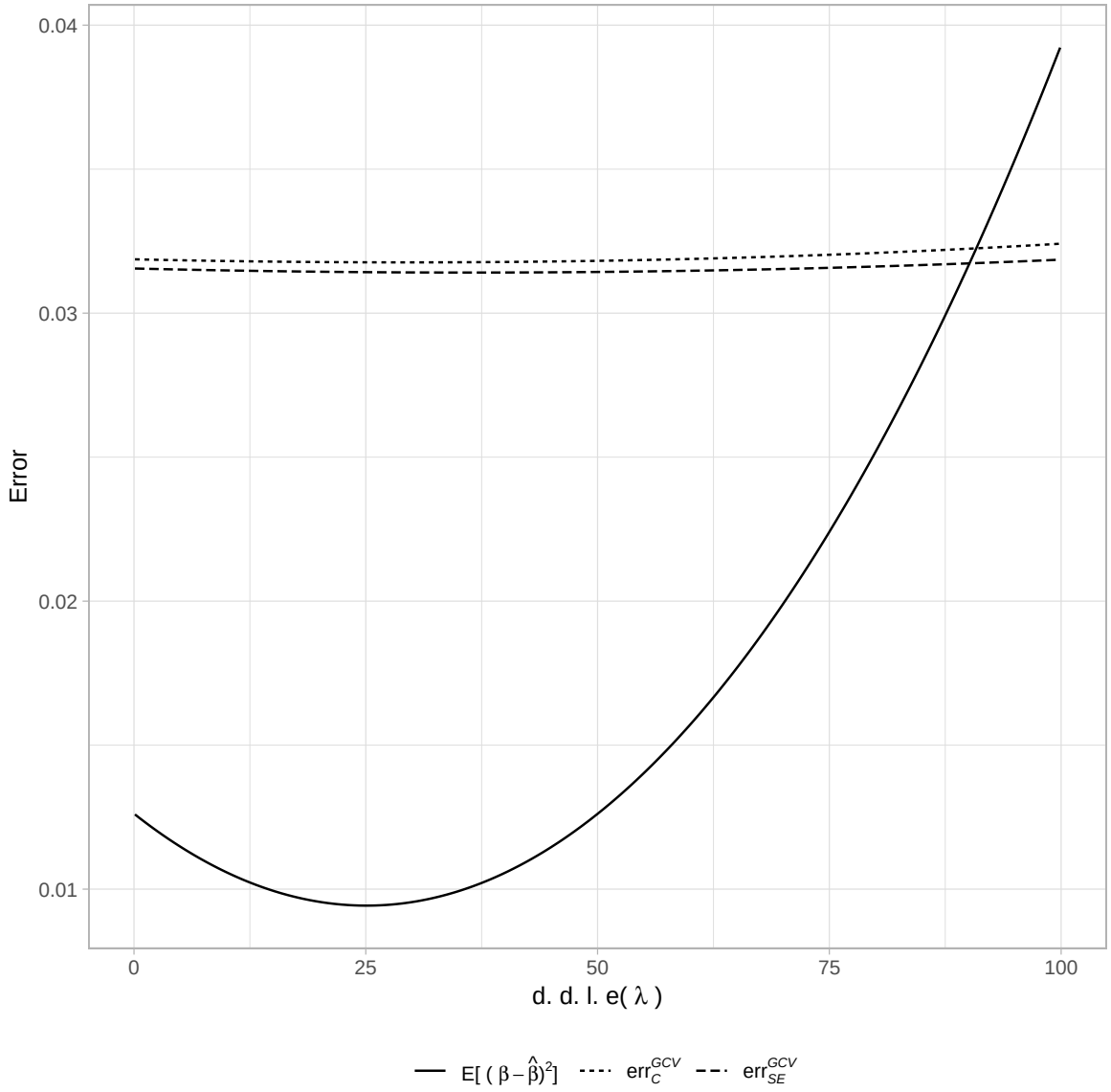


FIGURE 3.2 – Un exemple de GCV contrastée qui corrige le biais de la GCV en grande dimension et avec centrage empirique. La courbe pleine représente la précision de l'estimateur, la courbe en pointillés représente l'erreur de GCV et la courbe hachée représente l'erreur de GCV contrastée. En abscisse on utilise les degrés de liberté effectifs définis en (1.15) comme mesure de la complexité. L'étude est de taille n = 100 avec p = 2000.

Chapitre 4

Application de la régression ridge à l'estimation d'héritabilité

Dans cette section nous proposerons plusieurs idées pour regarder l'estimation d'héritabilité comme un problème d'apprentissage. Nous commencerons par présenter l'estimation d'héritabilité avec les modèles mixtes tel que proposé par Yang et al., puis proposerons plusieurs manières de quantifier l'héritabilité avec la régression ridge. Nous appliquerons ensuite nos idées sur des simulations, et proposerons également une application sur 4 phénotypes de la base de données UK-Biobank.

4.1 Notations et définitions des données

Dans cette section nous définirons les quantités que nous allons manipuler. Soit $\mathbf{M} \in \mathcal{M}_{n,p}(\mathbb{N})$ la matrice de génotypes avec codage additif de nos données i.e. $\forall i, j \in [\![1, n]\!] \times [\![1, p]\!],\ M_{i,j} \in \{0, 1, 2\}$. Nous supposerons que notre phénotype $\mathbf{y} \in \mathbb{R}^n$ suit le modèle polygénique additif

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} \tag{4.1}$$

$$= \mathbf{f} + \mathbf{g} + \mathbf{e}. \tag{4.2}$$

avec

- $\mathbf{f} = \mathbf{X}\beta$ un terme d'effets confondants non-génétiques avec $\mathbf{X} \in \mathcal{M}_{n,r}(\mathbb{R})$ et $\beta \in \mathbb{R}^r$. On trouvera dans ce terme l'intercept mais également des variables cliniques telles que le sexe ou l'âge.
- $\mathbf{g} = \mathbf{Z}u$ le terme d'effets génétiques avec $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$, la version standardisée de la matrice de génotypes $\mathbf{M}$, et $u \in \mathbb{R}^p$ le vecteur des effets génétiques.

— $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ un vecteur de bruit appelé terme d'environnement.

Rappelons que dans le cadre des GWAS nous avons $n << p$ et que nous sommes donc dans le cas des statistiques en grande dimension. L'estimation de l'héritabilité d'un point de vue apprentissage soulève plusieurs questions et problèmes :

— L'héritabilité est définie comme un ratio de variance. Nous pourrions donc être tentés d'estimer les termes du modèle (4.2) puis de calculer leurs variances empiriques et d'en déduire une estimation de l'héritabilité. Toutefois dans notre contexte de grande dimension la régression des moindres carrés n'est pas applicable pour estimer u et β. Il faudra donc utiliser d'autres méthodes linéaires.

— Un autre questionnement est sur quel ensemble estimer les variances ? Supposant que l'estimation soit réalisable et que nous disposions de $\hat{u}$ et d'un $\hat{\beta}$ estimés sur un ensemble d'apprentissage. Devons-nous effectuer l'estimation de $\text{var}(\hat{\mathbf{g}})$ (la variance des composantes du vecteur $\hat{\mathbf{g}}$) sur l'ensemble d'apprentissage et donc les mêmes données que celles servant à la construction de $\hat{u}$, ou bien sur un ensemble de test indépendant ? La définition d'héritabilité est indépendante du concept de surapprentissage mais nous savons que l'absence d'ensemble de test va mener à du surapprentissage.

— Enfin une dernière question se pose sur l'intégration des effets fixes. En effet l'héritabilité est définie comme la part de variance phénotypique due à la génétique, mais dans la littérature la place des effets fixes n'est pas forcement toujours bien définie.

4.2 L'estimation d'héritabilité par les modèles mixtes

L'équipe de Visscher propose d'utiliser les *genomic best linear unbiased predictors* (GBLUP) sur des individus non-apparentés pour l'estimation d'héritabilité [Yang et al., 2010]. Les GBLUP sont des BLUP utilisant une matrice d'apparentement issue de la matrice des marqueurs génétiques et correspondent au cas que nous avons présenté dans 2.6. Ces modèles sont largement utilisés en génétique animale et végétale mais à notre connaissance l'équipe de Visscher fut la première à proposer leur utilisation pour l'estimation d'héritabilité en génétique humaine.

Supposons que le phénotype suit un modèle polygénique linéaire : en utilisant les

mêmes notations que plus haut, nous pouvons poser

$$\mathbf{y} = \mathbf{Z}_* u + \mathbf{e} = \mathbf{g}_* + \mathbf{e} \tag{4.3}$$

avec $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, $\mathbf{Z}_* \in \mathcal{M}_{n,p_*}(\mathbb{R})$ la matrice de génotypes standardisés des variants causaux associés au phénotype. Une particularité est que dans la modélisation on suppose $u \sim \mathcal{N}(0_{p_*}, \tau \mathbf{I}_{p_*})$ i.e. nous travaillons avec un modèle à effets aléatoires. Nous appellerons ce modèle "modèle de Visscher". Nous remarquons en particulier que ce modèle suppose que tous les variants du modèle sont causaux. Nous supposerons également que les termes $\mathbf{g}$ et $\mathbf{e}$ sont indépendants :

$$\begin{aligned} \mathbf{g}_* &\sim \mathcal{N}(\mathbf{0}_n, \tau \mathbf{Z}_* \mathbf{Z}_*^T), \\ \mathbf{y} &\sim \mathcal{N}(\mathbf{0}_n, \tau \mathbf{Z}_* \mathbf{Z}_*^T + \sigma^2 \mathbf{I}_n). \end{aligned}$$

Si nous supposons que les individus sont non apparentés et que les variants sont indépendants, nous pouvons estimer la variance de chacune des composantes du terme génétique par

$$\mathrm{var}([g_*]_i) = \sigma_g^2 \simeq p_* \tau.$$

Nous arrivons à la forme de la matrice de covariance de $\mathbf{y}$

$$\begin{aligned} \mathrm{var}(\mathbf{y}) &= \tau \mathbf{Z}_* \mathbf{Z}_*^T + \sigma^2 \mathbf{I}_n = p_* \tau \times \frac{1}{p_*} \mathbf{Z}_* \mathbf{Z}_*^T + \sigma^2 \mathbf{I}_n \\ &= \sigma_g^2 \mathbf{G}_* + \sigma^2 \mathbf{I}_n. \end{aligned}$$

avec $\mathbf{G}_*$ la matrice de ressemblance génétique entre paires d'individus pour les variants causaux. Cette forme de matrice de covariance ressemble dans l'écriture à celle des calculs originaux de Fisher [1919] (même si elle est très différente puisque l'aléatoire n'est pas sur les génotypes mais les effets) et permet d'approcher l'héritabilité par la proportion de variance phénotypique due aux variants causaux

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2} = \frac{p_* \tau}{p_* \tau + \sigma^2}. \tag{4.4}$$

Cette quantité est parfois appelée héritabilité génomique et est une approximation de l'héritabilité au sens faible (que nous avons définie dans l'introduction).

En pratique nous ne savons que peu de choses sur l'emplacement et l'effet des

variants causaux. Nous ne pouvons donc pas construire $\mathbf{G}_*$ et n'avons pas accès à τ et σ^2. Visscher suggère d'approximer cette matrice $\mathbf{G}_*$ par une version "empirique" calculée sur la matrice des marqueurs dont on dispose (appelée *Genetic Relationship Matrix* et abrégée en *GRM*). Puisque nous supposons les individus non-apparentés, nous pouvons voir la GRM comme une matrice capturant les ressemblances cryptiques entre individus. Cela revient à poser un modèle à effets aléatoires utilisant $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ la matrice des variants génotypés de notre étude comme matrice de données :

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e} = \mathbf{g} + \mathbf{e}.$$

Nous estimons τ et σ^2 avec des vraisemblances restreintes et enfin l'héritabilité génomique par

$$\hat{h}^2_G = \frac{p\hat{\tau}}{p\hat{\tau} + \hat{\sigma}^2}.$$

Une partie du signal ne peut être captée car les variants causaux sont en déséquilibre de liaison incomplet (abrégé LD) avec les marqueurs génotypés. Yang et al. [2010] propose une correction pour la matrice $\mathbf{G}$ pour prendre en compte ce DL incomplet.

Utiliser un modèle à effets aléatoires sur la matrice des variants génotypés a plusieurs avantages :

— Elle permet d'écrire la matrice de covariance du phénotype sous la même forme que dans les modèles de Fisher et nous pouvons alors en extraire une approximation de l'héritabilité.

— Comme elle utilise des individus non-apparentés, la méthode se veut non-biaisée car échappant à l'environnement familial partagé.

— L'utilisation des modèles mixtes est en accord avec l'hypothèse d'une multitude d'effets faibles répartis sur tout le génome.

4.3 Régression ridge et héritabilités

Un des objectifs de cette thèse était d'essayer de faire rencontrer le concept d'héritabilité et l'apprentissage statistique. Dans cette section nous présenterons nos tentatives pour marier ces concepts, avec la régression ridge comme fil commun. Les quantités que

nous présenterons n'auront pas nécessairement de lien direct avec l'héritabilité définie par Fisher mais essayeront d'intégrer le concept de qualité d'apprentissage. Les quantités présentées dans cette section seront évaluées sur des simulations et des données réelles dans les sections suivantes.

4.3.1 Lien entre h^2 et λ

Il existe un lien direct entre paramètre de pénalisation de la régression ridge et héritabilité si nous acceptons le modèle de Visscher comme vérité et supposons que nous avons accès aux variants causaux. Nous rappelons qu'il existe un lien entre la régression ridge et les modèles à effets aléatoires 2.6.1. Grâce à ce lien nous avons en particulier $\lambda = \frac{\sigma^2}{\tau}$.

Avec un peu d'algèbre nous pouvons écrire l'approximation de l'héritabilité définie en 4.4 comme une fonction de λ [de Vlaming and Groenen, 2015]

$$h^2_G = \frac{p_*}{p_* + \lambda} \ ; \ \lambda = p_* \frac{1 - h^2_G}{h^2_G}. \tag{4.5}$$

En pratique nous n'avons évidemment toujours pas accès aux variants causaux mais nous pouvons écrire cette approximation avec les marqueurs. Nous commencerons par estimer le paramètre de pénalisation optimal $\hat{\lambda}_{opt}$ à partir de notre matrice de marqueurs $\mathbf{Z}$ et de $\mathbf{y}$ puis nous pourrons en déduire une estimation de l'héritabilité

$$\hat{h}^2_G = \frac{p}{p + \hat{\lambda}_{opt}} \ ; \ \hat{\lambda}_{opt} = p \frac{1 - \hat{h}^2_G}{\hat{h}^2_G}. \tag{4.6}$$

En choisissant λ qui minimise un critère de sélection nous aurons un premier lien entre calcul d'héritabilité et régression ridge.

Ce lien permet également de résoudre une des problématiques de la régression ridge : quelles sont les valeurs de paramètre de pénalisation pertinentes pour chercher λ_{opt} ? Pour les méthodes basées sur la cross-validation que nous avons présentées dans la section 2.5.2, nous choisissons une grille de λ sur laquelle on calculera une erreur pour approcher λ_{opt}. Comme la seule contrainte imposée à λ est d'être positif, le choix de cette grille n'est pas trivial. Toutefois en utilisant le lien 4.6 nous pouvons facilement construire une grille de λ associée à une grille d'héritabilité.

$$\underbrace{\{0.01, 0.02, ..., \ 0.99\}}_{h^2_g} \rightarrow \underbrace{\{p\frac{1-0.01}{0.01}, p\frac{1-0.02}{0.02}, ..., p\frac{1-0.99}{0.99}\}}_{\lambda}$$

Notons que cette approche pour choisir un paramètre de pénalisation optimal s'exporte hors du cadre de l'héritabilité, l'héritabilité pouvant être vue comme la variance expliquée du modèle.

4.3.2 Héritabilité prédictive

Dans cette section nous présenterons l'héritabilité prédictive (h_p^2) , une quantité généraliste mesurant la qualité de prédiction d'une méthode d'apprentissage, telle que la régression ridge. Bien qu'elle ne corresponde pas à l'héritabilité au sens strict du terme, les deux quantités ne semblent pas complètement indépendantes.

Définition

Pour définir h_p^2 nous supposerons le modèle généraliste suivant :

$$\mathbf{y} = f(\mathbf{Z}) + \mathbf{e} \tag{4.7}$$

avec :

- $\mathbf{y} \in \mathbb{R}^n$ un vecteur de phénotype.
- $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ une matrice de génotypes (normalisée).
- f une fonction lien.
- $\mathbf{e} \in \mathbb{R}^n$ un vecteur de bruit (qu'on supposera gaussien, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$).

L'héritabilité prédictive est définie de la manière suivante :

$$h_p^2(\mathbf{y}, \mathbf{Z}, \hat{\mathbf{y}}) = 1 - \frac{\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) - \mathcal{L}_{min}}{\mathcal{L}_{max} - \mathcal{L}_{min}} \tag{4.8}$$

avec :

- $\hat{f}$ un estimateur et $\hat{\mathbf{y}} = \hat{f}(\mathbf{Z})$ un vecteur d'estimation associé.
- $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ l'espérance d'une fonction de coût (risque) entre $\mathbf{y}$ et $\hat{\mathbf{y}}$.
- $\mathcal{L}_{max}$ et $\mathcal{L}_{min}$ respectivement les "pire" et "meilleure" valeurs possibles de $\mathcal{L}$.

Une question naturelle est la définition de $\mathcal{L}_{max}$ et $\mathcal{L}_{min}$. Nous avons fait le choix de supposer que pour $\mathcal{L}_{max}$, la réponse et les données sont indépendantes. A l'inverse, on suppose que le modèle n'est pas bruité pour le calcul de $\mathcal{L}_{min}$.

Nous n'avons pas défini si les prédictions étaient réalisées sur un ensemble de test indépendant ou sur l'ensemble d'apprentissage. Il est probable que les prédictions sans

ensemble de test seront trop optimistes mais le concept d'ensemble de test n'est pas défini dans l'héritabilité : nous calculerons h_p^2 avec et sans ensemble de test sur nos simulations.

Deux exemples d'h_p^2

Pour un phénotype quantitatif

Si nous supposons le phénotype quantitatif, un choix naturel pour $\mathcal{L}$ serait de prendre le risque quadratique.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \mathbb{E}\left[\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{y} - \hat{f}(\mathbf{Z})\right\|_2^2\right].$$

Sous l'hypothèse que, dans un cas idéal pour l'apprentissage, le modèle 4.7 soit non bruité, choisir $\mathcal{L}_{min} = 0$ est complètement intuitif.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \mathbb{E}\left[\left\|\mathbf{y} - \hat{f}(\mathbf{Z})\right\|_2^2\right] = \mathbb{E}\left[\|\mathbf{y}\|_2^2\right] - 2\mathbb{E}\left[\mathbf{y}^T \hat{f}(\mathbf{Z})\right] + \mathbb{E}\left[\left\|\hat{f}(\mathbf{Z})\right\|_2^2\right].$$

Sous l'hypothèse d'indépendance entre prédiction et réponse nous avons

$$\mathcal{L}_{max} = \mathbb{E}\left[\|\mathbf{y}\|_2^2\right] + \mathbb{E}\left[\left\|\hat{f}(\mathbf{Z})\right\|_2^2\right] = \mathrm{var}(\mathbf{y}) + \|\mathbb{E}[\mathbf{y}]\|_2^2 + \mathrm{var}\left(\hat{f}(\mathbf{Z})\right) + \left\|\mathbb{E}\left[\hat{f}(\mathbf{Z})\right]\right\|_2^2.$$

Sous l'hypothèse $\mathbb{E}[f(\mathbf{Z})] = 0$ et $\mathbb{E}[\mathbf{e}] = 0$, on a $\mathbb{E}[\mathbf{y}] = 0$. Nous devons alors choisir un $\hat{f}$ sous hypothèse d'indépendance. Un choix que nous pouvons faire est $\hat{f}(Z) = 0$ (i.e. l'estimateur nul) : nous prendrions comme "pire estimateur" celui qui quelles que soient les données ne renvoie que 0. C'est un choix relativement naturel, mais cela ne correspond pas au pire choix possible : si nous prenions $\hat{f}(\mathbf{Z}) = K$ avec K très grand, alors le risque quadratique serait également très élevé. En fait $\hat{f}(Z) = 0$ correspond au meilleur estimateur possible sous l'hypothèse d'indépendance.

En supposant $\hat{f}(\mathbf{Z}) = 0$, nous avons donc

$$\mathcal{L}_{max} = \mathrm{var}(\mathbf{y})$$

Nous n'avons pas accès au risque quadratique $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ et à la variance de $\mathbf{y}$, ils sont donc estimés par l'estimateur non biaisé de la variance :

$$\hat{\mathcal{L}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n-1} \left\| \mathbf{y} - \hat{f}(\mathbf{Z}) \right\|_2^2$$
$$\hat{\text{var}}(\mathbf{y}) = \frac{1}{n-1} \|\mathbf{y}\|_2^2$$

Et nous avons donc

$$h_p^2 = 1 - \frac{\left\| \mathbf{y} - \hat{f}(\mathbf{Z}) \right\|_2^2}{\|\mathbf{y}\|_2^2} = R^2(\mathbf{y}, \hat{\mathbf{y}}) \tag{4.9}$$

Nous retrouvons le coefficient de détermination R^2 usuel en statistique.

Pour un phénotype binaire

Si nous supposons le phénotype qualitatif, un choix naturel pour $\mathcal{L}$ serait de prendre le coût 0-1.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \mathbb{E}\left[\mathbb{1}_{\mathbf{y} \neq \hat{f}(\mathbf{Z})}\right] = \text{P}\left(\mathbf{y} \neq \hat{f}(\mathbf{Z})\right)$$

Sous l'hypothèse que le modèle 4.7 est non bruité, choisir $\mathcal{L}_{min} = 0$ est complètement intuitif.

Avant de définir $\mathcal{L}_{max}$, rappelons que

$$\begin{aligned}\mathcal{L} = \text{P}\left(\mathbf{y} \neq \hat{f}(\mathbf{Z})\right) &= \text{P}\left(\mathbf{y} \neq \hat{f}(\mathbf{Z}) \mid \mathbf{y} = 0\right) \text{P}(\mathbf{y} = 0) + \text{P}\left(\mathbf{y} \neq \hat{f}(\mathbf{Z}) \mid \mathbf{y} = 1\right) \text{P}(\mathbf{y} = 1) \\ &= \text{P}\left(\hat{f}(\mathbf{Z}) = 1\right) \text{P}(\mathbf{y} = 0) + \text{P}\left(\hat{f}(\mathbf{Z}) = 0\right) \text{P}(\mathbf{y} = 1)\end{aligned}$$

Notons K la prévalence de $\mathbf{y}$ et prenons $\hat{f}$ un estimateur aléatoire de prévalence K_f. Nous avons bien l'indépendance entre les données et cela semble correspondre au "pire cas possible".

$$\begin{aligned}\mathcal{L}_{max} &= \text{P}\left(\hat{f}(\mathbf{Z}) = 1\right) \text{P}(\mathbf{y} = 0) + \text{P}\left(\hat{f}(\mathbf{Z}) = 0\right) \text{P}(\mathbf{y} = 1) \\ &= K_f(1 - K) + (1 - K_f) K\end{aligned}$$

Cette valeur de $\mathcal{L}_{max}$ correspond au "no-information criterion" comme défini dans [Ambroise and McLachlan, 2002].

4.3.3 L'héritabilité comme un ratio de variances

Pour cette dernière manière d'approcher l'héritabilité nous allons retourner à sa définition i.e. à un ratio de variance. En notant (abusivement) var($\mathbf{x}$) la variance des composantes du vecteur $\mathbf{x}$ nous aurions

$$h^2 = \frac{\text{var}(\mathbf{g})}{\text{var}(\mathbf{y})}. \tag{4.10}$$

En pratique le terme $\mathbf{g}$ est inconnu, mais nous pouvons essayer de l'estimer (avec par exemple la régression ridge ou les modèles aléatoires). Nous définissons l'héritabilité ratio de variance empirique (h_r^2) comme

$$\hat{h}_r^2 = \frac{\text{var}(\hat{\mathbf{y}})}{\text{var}(\mathbf{y})} \tag{4.11}$$

avec $\hat{\mathbf{y}}$ un vecteur de prédiction. Cette quantité se justifie par sa ressemblance avec la définition originelle de l'héritabilité. Elle repose toutefois sur l'hypothèse que les prédictions que l'on va faire sont fonction de l'héritabilité. Comme pour l'héritabilité prédictive, le choix de l'ensemble sur lequel faire des prédictions ($\mathcal{A}$ ou $\mathcal{T}$) se pose. Nous testerons l'estimation de h_r^2 pour les deux ensembles sur des simulations et en utilisant la régression ridge comme méthode de prédiction.

4.4 Simulations et comparaisons avec l'approche REML

L'objectif de cette section est d'illustrer la capacité des quantités que nous avons définies dans la section précédente à capturer l'héritabilité avec des simulations. Nous utiliserons le lien entre héritabilité et paramètre de pénalisation, l'héritabilité prédictive (qui sera ici confondue avec le coefficient de détermination R^2) et l'héritabilité de ratio. L'héritabilité prédictive et l'héritabilité de ratio utiliseront la régression ridge pour estimer le vecteur d'effet génétique car la régression ridge nous permet de comprendre le lien entre prédiction et héritabilité grâce au point de vue de l'apprentissage statistique. Nous utiliserons la GCV pour choisir le paramètre de pénalisation (ce qui nous permettra de vérifier si elle estime bien le paramètre de pénalisation optimal en grande dimension). Les prédictions seront réalisées sur l'ensemble d'apprentissage ainsi que sur un ensemble de test indépendant.

4.4.1 Descriptif des simulations

Nous allons tester les quantités définies dans la section précédente à l'aide de simulations. Notre modèle de simulations est grandement inspiré de celui de Golan et al. [2014] et de Vlaming and Groenen [2015].

Dans ces simulations nous souhaitons évaluer l'influence de plusieurs paramètres : les dimensions de l'étude, l'héritabilité simulée et la proportion de variants causaux dans le modèle. Les différents niveaux de paramètres que nous allons tester sont résumés dans la table 4.1.

Paramètres	**Niveaux**
(n,p)	Génotypes simulés : (1000,10000) ; (5000,100000) ; (10000,500000) Génotypes de UKBB : (1000,10000) ; (5000,100000) ; (10000,417106)
f_c	0.1 ; 0.5 ; 1
h^2_{sim}	$\{0.1, ..., 0.9\}$

TABLEAU 4.1 – Table des valeurs des paramètres pour les simulations. n/p : ratio des dimensions de la matrices de marqueurs. f_c : proportion de variants causaux. h^2_{sim} : héritabilité simulée.

Nous allons proposer deux types de simulation selon la nature de la matrice de génotypes : dans les simulations que nous appellerons "synthétiques" les matrices de génotypes seront intégralement simulées tandis que dans les simulations "semi-synthétiques" ces matrices seront extraites des données de UK Biobank (que nous décrirons plus en détail dans la section 4.5).

Pour chaque valeur de (n, p), f_c et h^2_{sim} dans 4.1 nous découpons notre matrice de génotypes non standardisés en 3 blocs indépendants

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\mathcal{A}} \\ \mathbf{M}_{\mathcal{S}} \\ \mathbf{M}_{\mathcal{T}} \end{bmatrix} \in \mathcal{M}_{n+n_{\mathcal{S}}+n_{\mathcal{T}},p}.$$

avec $\mathbf{M}_{\mathcal{A}}$ la matrice pour l'apprentissage de l'estimateur, $\mathbf{M}_{\mathcal{T}}$ la matrice pour l'ensemble de test et enfin $\mathbf{M}_{\mathcal{S}}$ un ensemble de données indépendant que nous utiliserons pour corriger la GCV tel que décrit en 3.3.2.

Les matrices de génotypes des simulations synthétiques sont simulées de la manière suivante : nous commençons par simuler f un vecteur de fréquences alléliques de taille $p_{max} = 50\ 000$ dont les composantes suivent une loi uniforme $f \sim \mathcal{U}_p(0.05, 0.5)$, puis nous simulons les génotypes que l'on concatène pour former $\mathbf{M}_{\mathcal{A}}$, $\mathbf{M}_{\mathcal{S}}$ et $\mathbf{M}_{\mathcal{T}}$. Enfin

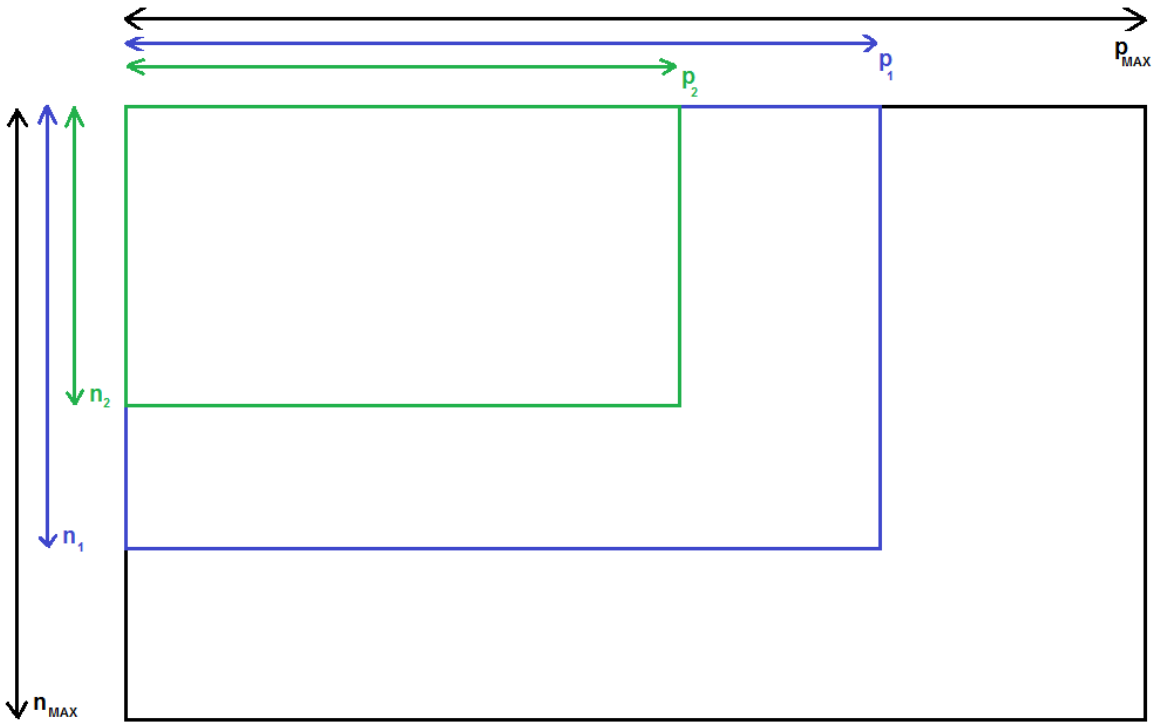


FIGURE 4.1 – Découpage de la matrice de données de l'ensemble d'apprentissage. Les matrices de génotypes incluent $(n_{max}, n_1, n_2) = (10000, 5000, 1000)$ individus. Pour les génotypes simulés $(p_{max}, p_1, p_2) = (5 \times 10^5, 10^5, 10^4)$ et pour les génotypes de UK-Biobank $(p_{max}, p_1, p_2) = (417106, 10^5, 10^4)$

.

nous standardisons ces matrices en utilisant respectivement $2f$ et $\sqrt{2f(1-f)}$ comme vecteur de moyennes et d'écart-types pour obtenir $\mathbf{Z}_\mathcal{A}$, $\mathbf{Z}_\mathcal{S}$ et $\mathbf{Z}_\mathcal{T}$:

$$\forall k \in [\![1, n]\!],\ \forall \mathcal{E} \in \{\mathcal{A}, \mathcal{S}, \mathcal{T}\},\ [Z_\mathcal{E}]_k = ([M_\mathcal{E}]_k - 2f) \times \text{diag}\left(2f(1-f)\right)^{-1/2}$$

avec $[M_\mathcal{E}]_k$ la k-ème ligne de $M_\mathcal{E}$.

Pour les matrices de génotypes de l'approche semi-synthétique, nous extrayons 3 sous-échantillons d'individus de taille n, $n_\mathcal{T}$ et $n_\mathcal{S}$ correspondant respectivement à $\mathbf{M}_\mathcal{A}$, $\mathbf{M}_\mathcal{T}$ et $\mathbf{M}_\mathcal{S}$. Pour la standardisation, nous calculons f_{UKBB} le vecteur de fréquences alléliques des variants sur UKBiobank puis nous standardisons les matrices en utilisant respectivement $2f_{UKBB}$ et $\sqrt{2f_{UKBB}(1-f_{UKBB})}$ comme vecteurs de moyennes et d'écart-types.

La régression ridge va demander le calcul de l'ACP de la matrice $\mathbf{Z}_\mathcal{A}\mathbf{Z}_\mathcal{A}^T$, le produit matriciel de matrices de grande taille et diverses autres manipulations de matrices très coûteuses en temps de calcul. Nous allons utiliser l'idée de de Vlaming and Groenen

[2015] pour que ces temps de calcul restent raisonnables : au lieu de simuler des matrices de génotypes pour chacune des combinaisons de paramètres $(n,p)/f_c/h^2_{sim}$ nous "fixons" la matrice de génotypes et allons varier les vecteurs d'effets génétiques u et environnementaux e pour toutes les combinaisons de f_c et h^2_{sim}. Ainsi nous n'avons à calculer l'ACP qu'une fois au lieu de 3×9 fois, ce qui va très fortement réduire les temps de calcul. Pour diminuer encore les temps de calcul, nous ne générons qu'une unique matrice de génotypes de dimensions $n_{max} \times p_{max}$ que nous découperons pour chacun des couples (n,p) d'intérêt (voir 4.1).

Pour une taille d'étude n de l'ensemble d'apprentissage et un nombre de variants fixé p, les phénotypes sont alors générés de la manière suivante :

1. Calcul des nombres de variants causaux et non-causaux

$$p_{\text{causaux}} = f_c \times p,$$
$$p_{\text{non-causaux}} = p - p_{\text{causaux}}.$$

2. Génération du vecteur d'effets génétiques pour les variants causaux

$$u_{\text{causaux}}(p, f_c, h^2_{sim}) \sim \mathcal{N}\left(0_{p_{\text{causaux}}}, \frac{h^2_{sim}}{p_{\text{causaux}}}\mathbf{I}_{p_{\text{causaux}}}\right),$$

puis construction du vecteur d'effets génétiques pour tous les variants

$$u(p, f_c, h^2_{sim}) = \begin{pmatrix} u_{\text{causaux}}(p, f_c, h^2_{sim}) \\ 0_{p_{\text{non-causaux}}} \end{pmatrix}.$$

3. Pour chaque ensemble $\mathcal{E} \in \{\mathcal{A}, \mathcal{S}, \mathcal{T}\}$ (de cardinal $n_{\mathcal{E}}$),
   (a) Création des termes génétiques

$$\mathbf{g}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim}) = \mathbf{Z}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim})\ u(n_{\mathcal{E}}, p, f_c, h^2_{sim}).$$

   (b) Création du terme d'environnement

$$\mathbf{e}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim}) \sim \mathcal{N}(\mathbf{0}_{n_{\mathcal{E}}}, (1 - h^2_{sim})\mathbf{I}_n).$$

   (c) Concaténation pour former le phénotype

$$\mathbf{y}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim}) = \mathbf{g}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim}) + \mathbf{e}_{\mathcal{E}}(n_{\mathcal{E}}, p, f_c, h^2_{sim}).$$

4.4.2 Analyse de l'estimation d'h_g^2

Présentation des résultats

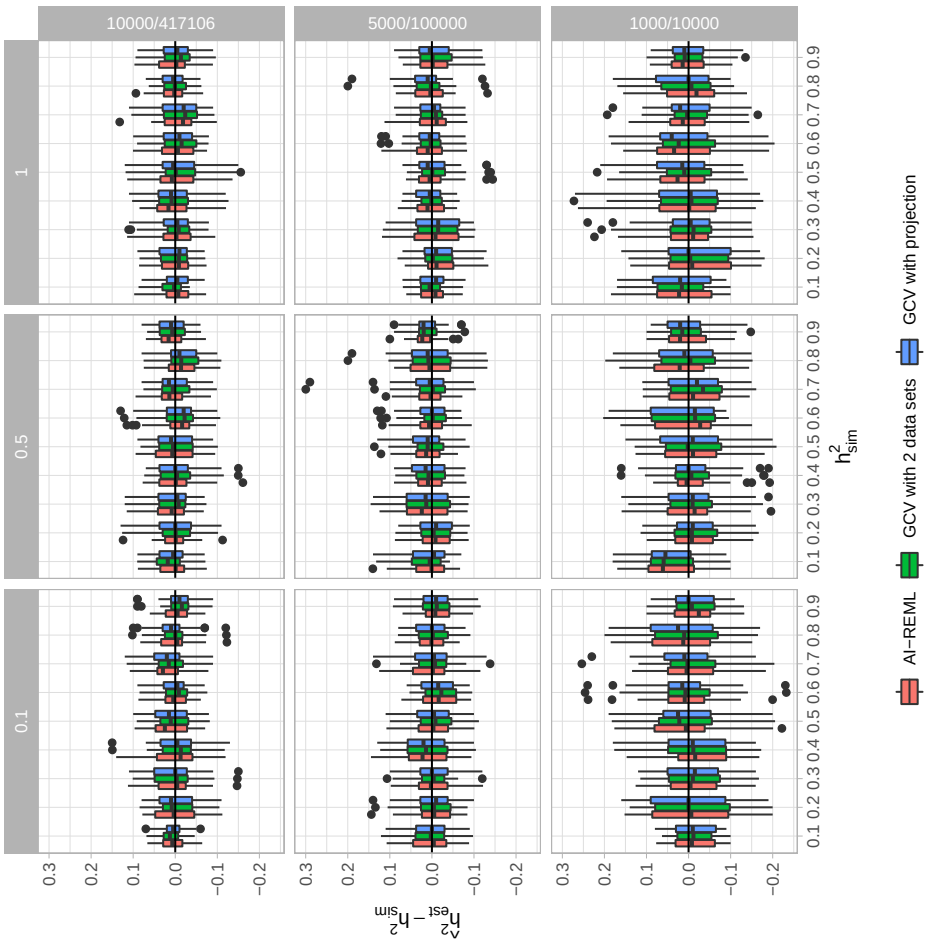
Les résultats d'estimation d'h^2 pour les simulations synthétiques et semi-synthétiques sont présentés dans la figure 4.2. La sous-figure 4.2a représente les résultats pour les simulations synthétiques et la sous-figure 4.2b les résultats pour les simulations semi-synthétiques. Chacune de ces figures est composée d'une grille de 9 graphes. Les lignes correspondent aux différents ratios n/p et les colonnes à différents pourcentages de variants causaux f_c. Pour chacun de ces graphes l'axe des abscisses correspond à 9 valeurs d'héritabilité simulée et l'axe des ordonnées à la différence entre héritabilité estimée et héritabilité simulée. Les différents niveaux des paramètres n/p, f_c et h^2_{sim} sont synthétisés dans la table 4.1. Dans les deux approches de simulations nous estimons et représenterons dans un boxplot h^2 avec la régression ridge en utilisant la GCV pour obtenir le paramètre de pénalisation optimal puis le lien 4.6 pour avoir une estimation de l'héritabilité. Pour l'estimation d'héritabilité avec la GCV, nous regarderons les corrections avec une matrice de contraste et un deuxième jeu de données pour le calcul de l'intercept. Nous regardons également l'estimation d'héritabilité faite avec les modèles mixtes (en utilisant un algorithme AI-REML) et une validation croisée 10-fold (uniquement pour les simulations synthétiques).

Sur les simulations synthétiques nous voyons que pour toutes les combinaisons de f_c, n/p et h^2_{sim}, les différences entre héritabilités simulée et estimée sont localisées autour de 0. Aucune des méthodes d'estimation ne semble donner de résultats très différents des autres. Nous remarquons une augmentation de la variance des estimations selon le ratio n/p : la variance des estimations est maximale pour 1000/100000 et minimale pour 10000/500000. Nous ne remarquons pas d'effet du pourcentage de variance causaux ou de l'héritabilité simulée.

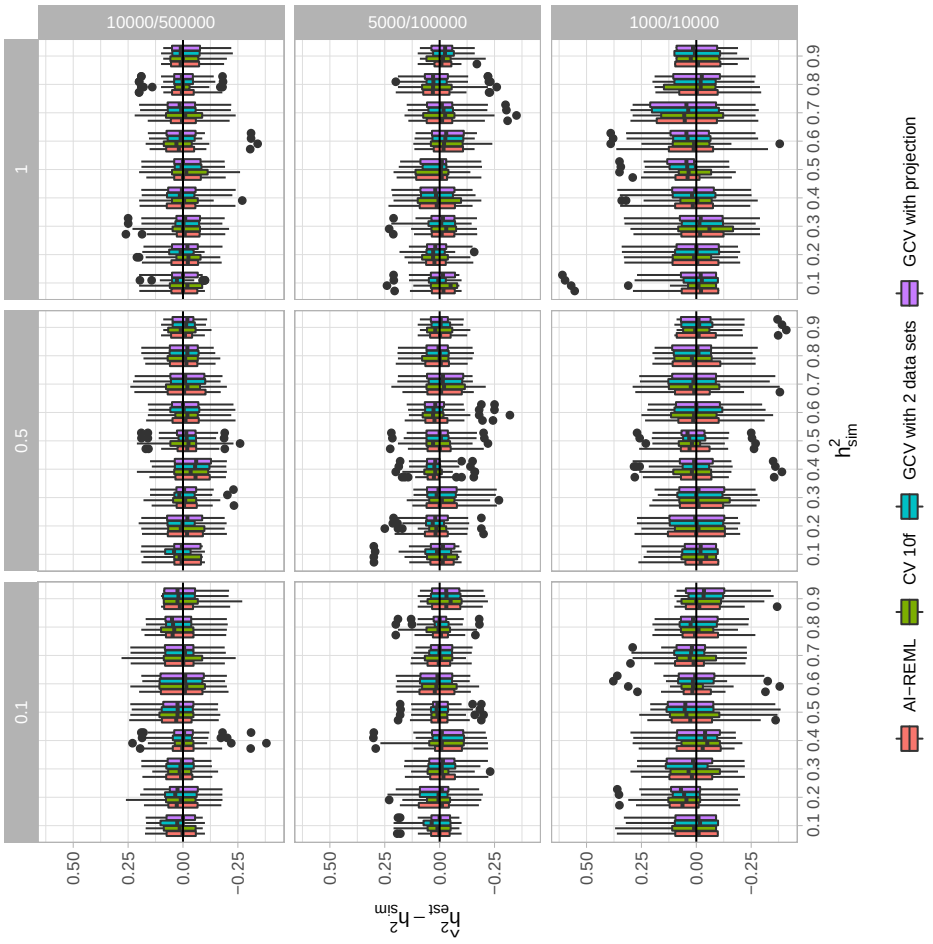
Les comportements selon n/p, f_c et h^2_{sim} sont les mêmes pour les simulations semi-synthétiques. Les différentes méthodes d'estimation donnent également des résultats similaires. Nous voyons également que la variance des estimations augmente quand le le ratio n/p diminue. Nous remarquons toutefois que la variance des estimations est plus faible que pour les simulations synthétiques à n/p équivalent.

Discussion

Nous pouvons tirer plusieurs conclusions de ces résultats. Commençons par ceux sur la GCV. Une première conclusion est que nos deux approches pour corriger la



(a) Simulation synthétique



(b) Simulation semi-synthétique

FIGURE 4.2 – Graphes d'estimation d'héritabilité génomique sur des simulations. La figure de gauche correspond aux simulations synthétiques et celle de droite aux simulations semi-synthétiques. Dans chaque figure on trouve 9 graphes rangés en colonne selon f_c et en ligne selon n/p. Chacun de ces graphes est un ensemble de boxplots dont l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à la différence entre l'héritabilité simulée et l'estimation d'héritabilité en utilisant le lien entre héritabilité et paramètre de pénalisation optimal. Les boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données, un algorithme AI-REML et (uniquement pour les simulations synthétiques) une régression ridge avec une validation croisée à 10 folds.

GCV semblent fonctionner pour les deux types de simulations. En effet la différence entre héritabilité simulée et estimée est très proche de 0 pour toutes les combinaisons de paramètres de simulations. Les deux corrections que nous avons proposées pour la GCV en grande dimension sont donc fonctionnelles et semblent donner des résultats équivalents.

Un autre constat que nous pouvons faire avec les simulations synthétiques est que la GCV donne des résultats proches de la validation croisée 10-fold. Ce résultat confirme la validité de la GCV corrigée en grande dimension.

En terme de calcul d'héritabilité, nous voyons qu'estimer l'héritabilité avec la régression ridge donne de bons résultats en moyenne. Nous voyons également que la variabilité des estimations semble du même ordre de grandeur que celles des méthodes basées sur le modèle mixte (utilisant ici l'AI-REML). Ceci confirme que la régression ridge peut être utilisée pour estimer l'héritabilité.

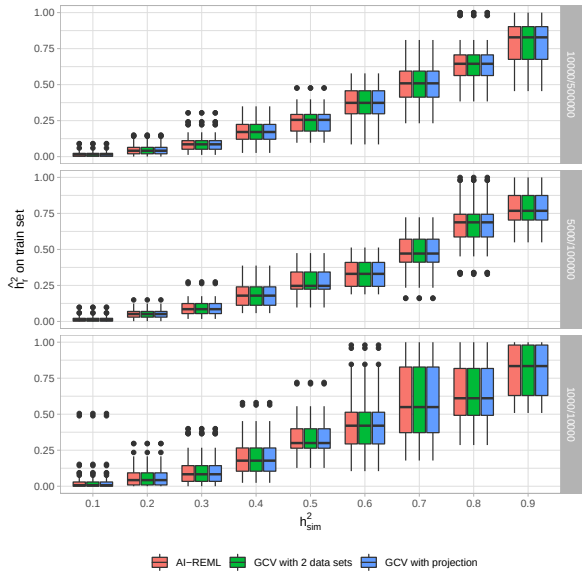
Le dernier constat porte sur la variance des estimations selon le ratio n/p. Au vu des résultats cette variance semble diminuer avec le ratio n/p, ce qui peut sembler contre intuitif. Cependant, d'après Visscher and Goddard [2015], la variance des estimations est censée diminuer quand n augmente ce qui sera le cas ici (n augmente mais le rapport n/p diminue). Toutefois nous remarquons que pour un ratio n/p équivalent la variance est plus faible pour les simulations semi-synthétiques. Une explication est que les variants de la matrice des génotypes de UKBiobank sont corrélés entre eux, ce qui diminue le nombre de variants effectifs du modèle. Nous pouvons donc en conclure que le nombre de variants importe aussi dans la variance des estimations.

4.4.3 Analyse de l'estimation de h_p^2 et h_r^2

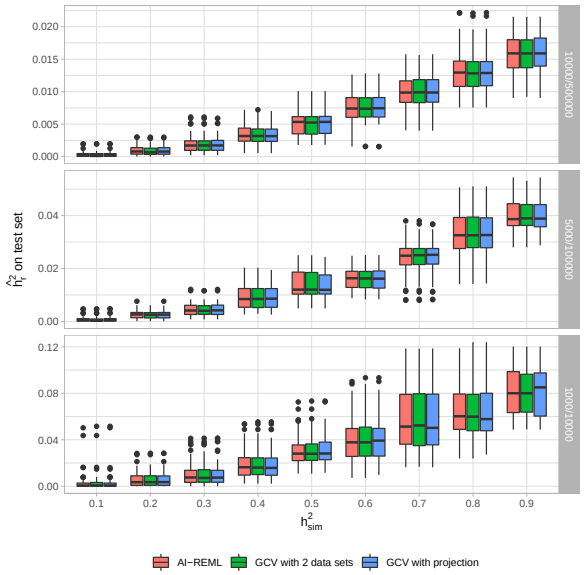
Présentation des résultats

Les estimations d'h_r^2 et d'h_p^2 (qui rappelons est confondu avec le R^2) sont respectivement présentées dans les figures 4.3 et 4.4. Dans chacune de ces figures on trouve quatre sous-graphes. Ils correspondent aux quantités calculées sur les ensembles d'apprentissage $\mathcal{A}$ et de test $\mathcal{T}$ pour chacune des deux approches de simulation.

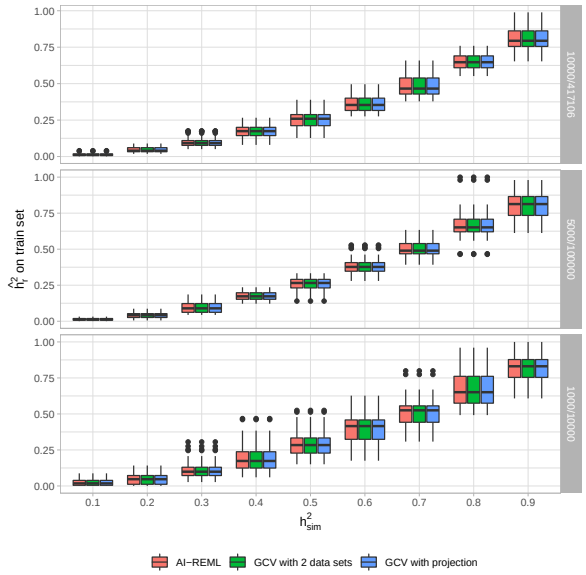
Dans chacune des sous-figures apparaissent respectivement l'estimation de la quantité selon l'héritabilité simulée pour les 3 ratios n/p présentés dans la table 4.1, pour chacune des simulations synthétiques et semi-synthétiques. Pour chaque étude nous avons calculé un estimateur sur l'ensemble d'apprentissage $\mathcal{A}$ avec la régression ridge en utilisant la GCV pour choisir le paramètre de pénalisation optimal. Nous évaluerons



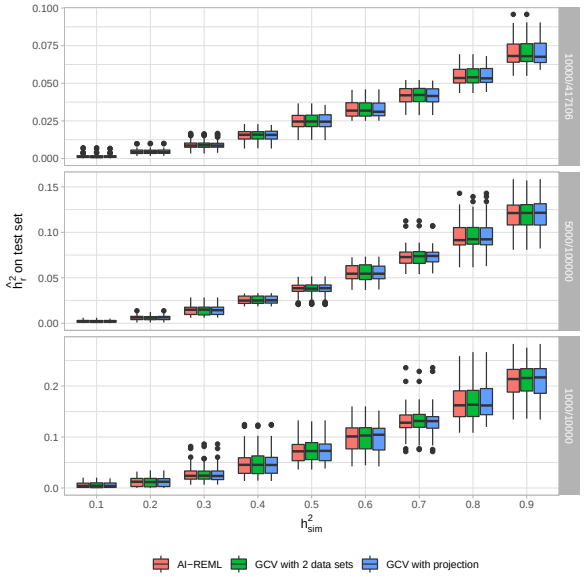
(a) h^2_r sur l'ensemble d'apprentissage, simulations synthétiques.



(b) h^2_r sur l'ensemble de test, simulations synthétiques.

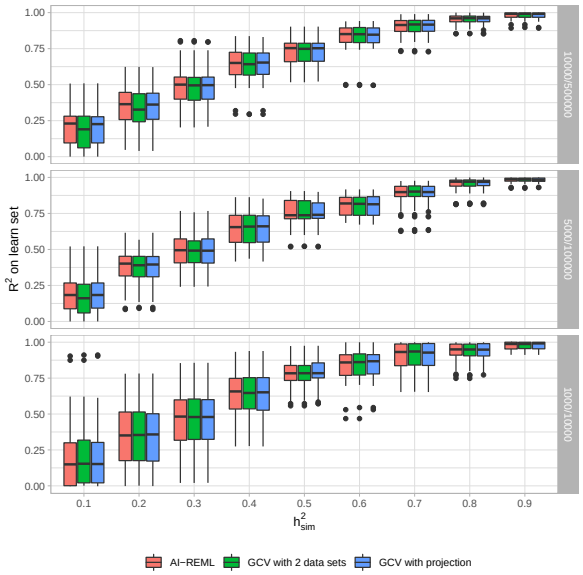


(c) h^2_r sur l'ensemble d'apprentissage, simulations semi-synthétiques.

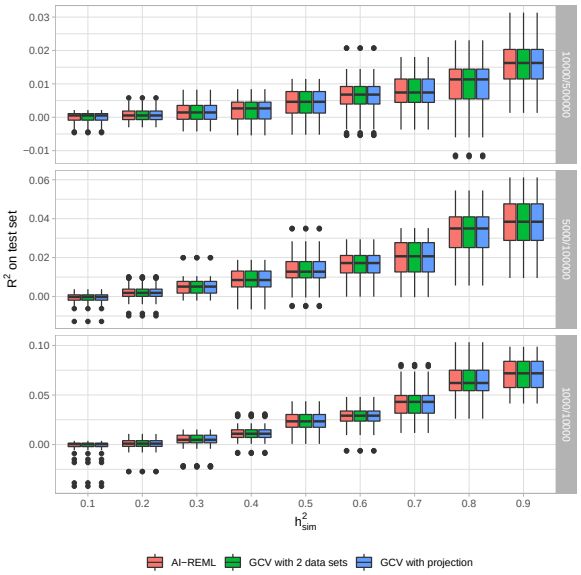


(d) h^2_r sur l'ensemble de test, simulations semi-synthétiques.

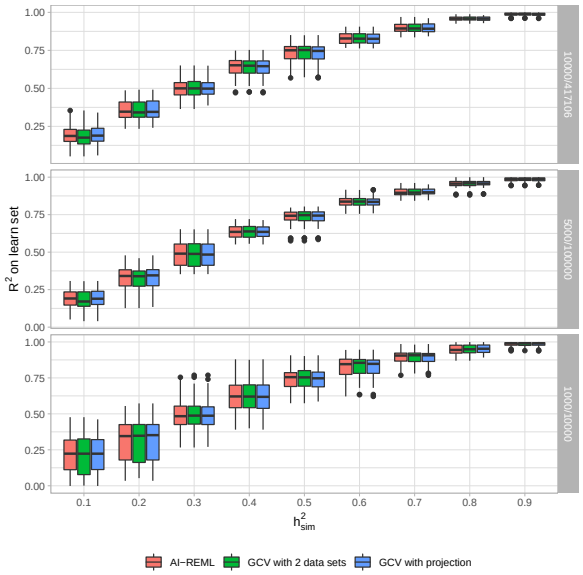
FIGURE 4.3 – Graphe d'estimation des h^2_r sur ensemble d'apprentissage et de test pour les simulations. Les figures du haut correspondent aux simulations synthétiques (celles du bas aux simulations semi-synthétiques) et les figures de gauche aux estimations sur $\mathcal{A}$ (celles de droite aux estimations sur $\mathcal{T}$). Dans chaque figure nous trouvons 3 graphes rangés en ligne selon n/p et dans chacun de ces graphes se trouvent des séries de boxplot. Pour chaque graphe l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h^2_r. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.



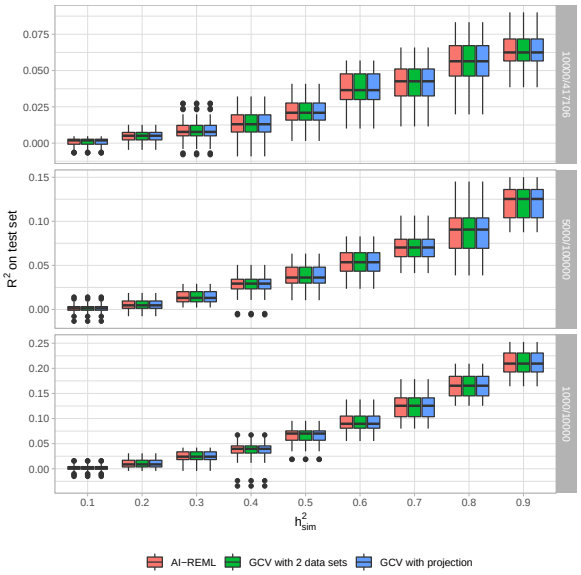
(a) h_p^2 sur l'ensemble d'apprentissage, simulations synthétiques.



(b) h_p^2 sur l'ensemble de test, simulations synthétiques.



(c) h_p^2 sur l'ensemble d'apprentissage, simulations semi-synthétiques.



(d) h_p^2 sur l'ensemble de test, simulations semi-synthétiques.

FIGURE 4.4 – Graphe d'estimation des h_p^2 (ou du R^2) sur ensemble d'apprentissage et de test pour les simulations. Les figures du haut correspondent aux simulations synthétiques (celles du bas aux simulations semi-synthétiques) et les figures de gauche aux estimations sur $\mathcal{A}$ (celles de droite aux estimations sur $\mathcal{T}$). Dans chaque figure nous trouvons 3 graphes rangés en ligne selon n/p et dans chacun de ces graphes se trouvent des séries de boxplot. Pour chaque graphe l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h_p^2. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.

les deux corrections de la GCV (avec une matrice de contraste et un second jeu de données) et calculerons également un BLUP issu du modèle mixte (en utilisant l'AI-REML pour l'estimation des paramètres de variance).

Notons que le paramètre f_c a disparu : pour alléger les figures nous présenterons uniquement les résultats pour $f_c = 0.1$ car nous n'avons pas observé d'effet de f_c sur les estimations dans la section d'avant. Les figures pour les autres valeurs de f_c sont disponibles en annexe B.

Commençons par regarder les résultats pour h_r^2 sur l'ensemble d'apprentissage. Nous observons le même comportement entre simulations synthétiques (figure 4.3a) et semi-synthétiques (figure 4.3c) : une augmentation (visiblement quadratique) de h_r^2 quand l'héritabilité simulée augmente. L'approche pour construire l'estimateur ne semble pas avoir d'importance. Les estimations sont du même ordre de grandeur que leur h_{sim}^2 associée mais toujours en dessous de l'héritabilité simulée. Notons que le ratio n/p ne semble pas avoir d'influence sur la moyenne de l'estimation. La dispersion des estimations semble être fonction de l'héritabilité simulée et du ratio n/p : elle semble croissante en h_{sim}^2 et croissante selon n/p. Nous remarquons également une plus grande variance des estimations pour les simulations synthétiques.

Regardons maintenant h_r^2 pour l'ensemble de test avec les graphiques 4.3b et 4.3d. Encore une fois, l'approche pour construire l'estimateur ne semble pas avoir d'importance et le comportement des estimations semble toujours être une fonction quadratique en h_{sim}^2. La grande différence avec l'estimation sur $\mathcal{A}$ est l'ordre de grandeur des estimations : il y a un fort biais de sous-estimation (la valeur maximale que nous observons pour les estimations sur $\mathcal{T}$ vaut environ 0.25 et vient des simulations semi-synthétiques avec $h_{sim}^2 = 0.9$ et $n/p = 10000/417106$). Nous observons également que les estimations sont en moyenne plus élevées sur les simulations semi-synthétiques. La variance des estimations se comporte comme pour l'ensemble d'apprentissage : elle croît en h_{sim}^2 et croît selon n/p .

D'après les figures 4.4a et 4.4c, les résultats sur le coefficient de détermination sur $\mathcal{A}$ présentent des similitudes avec ceux obtenus pour l'héritabilité de ratio. Nous observons une augmentation des quantités moyennes avec h_{sim}^2, la dispersion des estimations est bien plus grande pour les simulations synthétiques et quand n/p augmente, et encore une fois le choix de la méthode ne semble pas avoir d'importance. Nous notons tout de même des différences importantes entre h_p^2 et h_r^2 (toujours sur $\mathcal{A}$). En particulier nous surestimons l'héritabilité simulée et la variance des estimations diminue quand

h^2_{sim} augmente.

Pour les estimations sur $\mathcal{T}$ (figures 4.4b et 4.4d), nous retrouvons des résultats quasiment identiques entre h^2_p et h^2_r. La seule différence est que les valeurs de h^2_p peuvent être négatives au contraire de h^2_r qui est par définition positive. Sinon le comportement moyen, la variance des estimations et l'ordre de grandeur sont quasiment les mêmes pour les deux quantités.

Discussion

Le premier constat est que le choix de l'approche pour corriger la GCV n'a pas d'influence sur le pouvoir prédictif. C'est un résultat auquel nous nous attendions. Il faut toutefois noter que nos simulations n'avaient pas de covariables non-génétiques.

Pour les quantités calculées sur l'ensemble $\mathcal{T}$, le constat évident est qu'elles ne peuvent pas servir à estimer l'héritabilité en grande dimension : nous voyons en effet une très forte sous-estimation. Les prédictions étaient de bien meilleure qualité sur l'ensemble d'apprentissage et il y a donc (sans surprise) un fort surapprentissage. La sous-estimation est la plus forte pour les héritabilités simulées fortes et elle semble augmenter quand le ratio n/p est faible. Les résultats sont meilleurs pour les simulations semi-synthétiques mais restent très biaisés.

Ces résultats sur l'ensemble de test ne sont pas étonnants. En effet nous ne nous attendions pas à ce que la régression ridge ou les BLUP donnent d'excellents résultats en terme d'estimation des effets génétiques dans un contexte de grande dimension où les conditions d'apprentissage sont très mauvaises (et donc un très fort surapprentissage apparaît). Le fait que les résultats soient meilleurs pour les plus hautes valeurs de ratio n/p va dans ce sens puisqu'il s'agit des situations où les conditions pour l'apprentissage sont les moins mauvaises. Les meilleurs résultats des simulations semi-synthétiques par rapport aux simulations synthétiques vont dans le même sens : pour ces simulations semi-synthétiques les variants ne sont pas indépendants par construction et donc les variants corrélés diminuent le nombre de paramètres effectifs à apprendre. Nous sommes donc dans des meilleures conditions d'apprentissage d'où l'amélioration des capacités prédictives. Remarquons tout de même une amélioration des capacités prédictives quand h^2_{sim} augmente, ce qui nous rassure un peu : il y a un effet de l'héritabilité mais l'apprentissage est incomplet.

Les résultats sur l'ensemble d'apprentissage sont plus difficiles à analyser. Les deux quantités h^2_p et h^2_r augmentent avec h^2_{sim} et les estimations sont du même ordre de

grandeur que h^2_{sim}, toutefois nous avons noté une sous-estimation pour h^2_r et une sur-estimation pour h^2_p. Il semble donc que la variance du vecteur de prédiction $\hat{y}$ soit sous-estimée, ce qui entraîne mécaniquement un biais de sous-estimation pour h^2_r (de par sa formule). Nous savons que la grande dimension entraîne un grand terme de pénalité pour la ridge et donc un fort lissage des coefficients de l'estimateur ridge, ce qui pourrait expliquer la trop faible variance des coefficients du vecteur de prédiction. Nous confirmons cette intuition en voyant que l'h^2_r estime le mieux l'héritabilité quand elle est très forte (et que donc le paramètre de pénalisation de la régression ridge est au plus proche de 0) ou très faible (et que donc le paramètre de pénalisation est déjà naturellement très élevé).

Nous allons maintenant proposer une explication pour le fait que h^2_p soit trop optimiste. Nous savons que des prédictions sur l'ensemble d'apprentissage sont trop optimistes à cause du surapprentissage. Le bruit est donc sous-estimé et sa variance avec lui, ce qui va mécaniquement surestimer h^2_p (la formule du coefficient de détermination pouvant être vue comme 1 moins la variance du bruit estimée sur la variance de la réponse).

Au vu des résultats sur les deux quantités, il semblerait donc que dans ce contexte de grande dimension (défavorable à l'apprentissage), le phénomène de surapprentissage observé sur l'h^2_p lorsqu'elle est calculée sur $\mathcal{A}$ compense largement la faible variance des prédictions induite par la forte pénalisation.

Une grande différence par rapport aux calculs sur $\mathcal{T}$ est que ici le ratio n/p ne semble jouer un rôle que sur la variance des estimations (qui diminue quand le ratio diminue) et non pas sur le comportement moyen. Cette similarité sur le comportement moyen semble indiquer que le surapprentissage observable sur les prédictions de l'ensemble d'apprentissage ne semble plus impacté par le ratio n/p en grande dimension, comme si il avait atteint son maximum (nous proposerons une explication de ce maximum dans le chapitre suivant). De la même manière, la sous-estimation de la variance des prédictions semble aussi être maximale.

La diminution de la variance des estimations quand le ratio diminue est un résultat étonnant : nous nous serions plutôt attendus à ce que des conditions d'apprentissage plus difficiles (i.e. quand le ratio diminue) augmente la variance des estimations. Nous pouvons penser que ce résultat s'explique par le fait que la diminution du ratio est associée à une augmentation de la taille de $\mathcal{A}$ pour nos simulations. Le nombre de variants semble jouer un rôle puisque les simulations semi-synthétiques (qui grâce aux corrélations entre variants ont un plus petit nombre de paramètres effectif à apprendre)

présentent une dispersion plus faible des estimations mais il semble moins important que la taille de l'échantillon.

En conclusion, aucune de ces quantités ne semble donner une estimation vraiment satisfaisante de l'héritabilité dans un contexte de grande dimension. Nous reviendrons plus tard sur elles dans ce manuscrit pour étudier leur comportement en termes de pouvoir prédictif, ainsi que le comportement de la variance du vecteur d'estimation des effets.

4.5 Application aux données UKBiobank

Dans cette section nous allons essayer l'estimation d'héritabilité avec la régression ridge sur le jeu de données publiques UKBiobank. Le passage de données simulées à données réelles amène son lot de problématiques telles que le pré-traitement nécessaire des données, l'intégration des covariables non pénalisées et la prise en compte d'une éventuelle structure de population.

Nous nous intéresserons à l'estimation de l'héritabilité de 4 traits morphologiques : la taille, l'indice de masse corporelle (IMC), la circonférence des hanches et le tour de taille. Pour estimer l'héritabilité nous utiliserons le lien entre modèle à effets aléatoires et régression ridge (en utilisant la GCV corrigée par projection ou avec un deuxième jeu de données pour obtenir le paramètre de pénalisation), ainsi que l'estimation de l'héritabilité génomique avec des composantes de variances estimées par modèles à effets mixtes. Nous reviendrons sur les pouvoirs prédictifs dans un autre chapitre.

La manipulation des données et les filtres qualité ont été effectués avec R et en particulier le package `gaston` [Perdry and Dandine-Roulland, 2017]. Les calculs d'héritabilité étant en $\mathcal{O}(n^3)$, nous n'avons pas réussi à faire tourner nos algorithmes sur les 500 000 individus de UK Biobank (en particulier les divers produits matriciels et la SVD). Nous avons donc décidé de nous concentrer sur des sous-échantillons de 10 000 individus (et de 1 000 individus pour le second ensemble de l'approche à deux jeu de données) sur lesquels nous appliquerons les filtres puis estimerons l'héritabilité. Ce sous-échantillonnage sera répété 10 fois pour estimer la variabilité des estimations associée au jeu de données. Notons également que toujours pour des raisons calculatoires, nous inclurons dans ces analyses uniquement les variants génotypés et n'inclurons donc pas les variants imputés.

4.5.1 Description des données

L'étude UK Biobank est une étude prospective réalisée au Royaume-Uni [Sudlow et al., 2015]. Avec environ 500 000 patients, elle représente l'une des plus grandes bases de données disponible. Les patients âgés de 40 à 69 ans ont été recrutés et génotypés dans une vingtaine de centres dans le Royaume-Uni entre 2006 et 2010. Nous avons accès à des données mesurées venant de visites médicales et de questionnaires allant de la mesure de traits morphologiques à des maladies reportées par les participants. Les 50 000 premiers patients ont été génotypés avec la puce Affymetrix UK BiLEVE Axiom et les 450 000 suivants avec la puce Affymetrix UK Biobank Axiom®. Une des covariables fournies est l'origine ethnique auto-reportée des patients (la répartition des différentes ethnies est décrite dans la figure 4.5b).

En plus des données génétiques nous ajouterons comme covariables le sexe, l'année de naissance, le centre de recrutement, la puce utilisée pour le génotypage et des variables de structure de population.

L'analyse des composantes principales fournies par UK Biobank nous montre une grande diversité génétique au sein de l'étude (figure 4.5). Il est connu qu'une structure de population peut avoir un effet sur les estimations d'héritabilité. Pour prendre en compte cette potentielle structure, une technique très utilisée est d'effectuer une analyse en composantes principales (ACP) sur la matrice de covariance empirique des individus étudiés et de donner en covariables au modèle les composantes principales (ou les vecteurs propres) associées aux premières valeurs propres.

Nous avons décidés de regarder l'influence du choix de population et du choix de construction des covariables de structure. Nous allons travailler sur 3 populations :

— Dans le premier cas nous prenons nos individus aléatoirement sans aucun filtre (TTLM).

— Dans le second cas nous ne prenons que des individus reportés comme "White British" (WB).

— Enfin dans le dernier cas nous sélectionnons un sous-ensemble d'individus les plus représentatifs parmi les WB qu'on appellera "Verified White British" (VWB). Pour choisir ces individus nous avons tracé une ellipse au "centre" du graphe des composantes principales des individus WD, et les individus dans cette ellipse seront considérés comme les "Verified White British". L'équation de cette ellipse est donnée dans le graphique 4.5c.

Sur chacune de ces populations nous allons calculer les covariables de structure de 3 manières qui sont seront décrites dans une section suivante. Comme ces covariables doivent être calculées sur des données propres, nous allons commencer par décrire nos filtres qualité.

4.5.2 Description et influence des contrôles qualité

Nous décrivons ici les filtres de contrôle qualité pour les différentes approches et populations. Pour l'approche avec une matrice de projection, tous les filtres sont appliqués sur l'unique ensemble d'apprentissage. Pour l'approche à deux jeux de données, les filtres sur les individus sont appliqués sur les ensembles d'apprentissage et de standardisation tandis que les filtres sur les variants sont calculés uniquement sur l'ensemble de standardisation. Une fois ceux-ci terminés, nous listons les variants sélectionnés sur l'ensemble de standardisation et nous excluons tous les autres de l'ensemble d'apprentissage. Précisons que les seuils des filtres sont les mêmes pour les différentes populations.

Pour l'estimation d'héritabilité sur données de populations nous supposons que les individus sont indépendants les uns des autres. Pour s'assurer que nos individus ne sont pas apparentés et qu'ils ne partagent donc pas le même environnement, Yang et al. [2010] suggèrent d'identifier toutes les paires d'individus dont le coefficient de GRM associé est supérieur à 0.025 puis d'enlever un individu de chaque paire (en veillant à enlever en priorité ceux apparaissant dans plusieurs paires).

Voici la liste de nos filtres :

— Sélection des variants autosomiques.

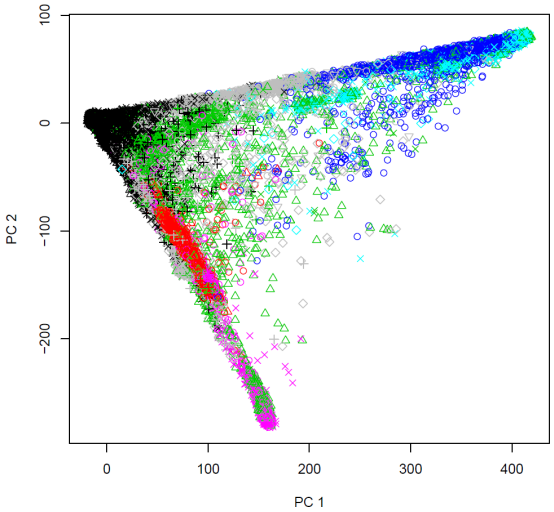
— Suppression des variants ET des individus avec un taux d'appel < 0.99.

— Suppression des variants avec une fréquence d'allèle mineur (MAF) < 0.01.

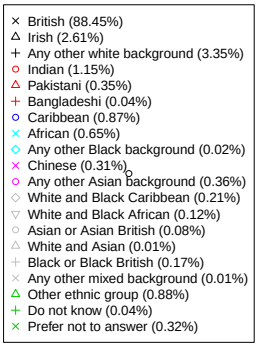
— Suppression de tous les variants non biallélіques.

— Suppression des variants avec une p-valeur $< 10^{-7}$ au test d'équilibre de Hardy-Weinberg.

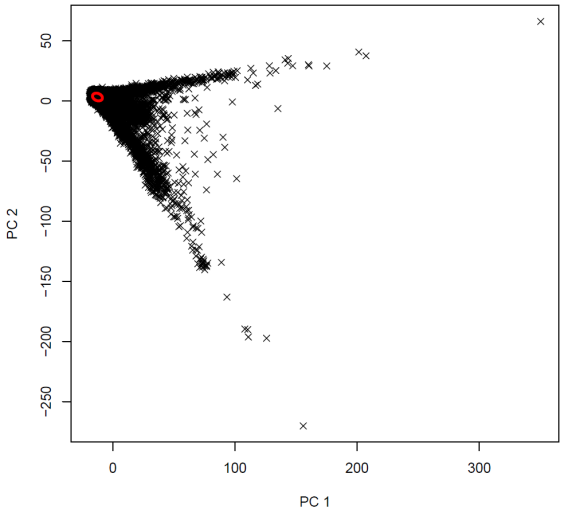
Les détails des résultats de ces contrôles qualité sont détaillés dans la table 4.2. Nous remarquons que les résultats finaux sont assez proches pour les différentes populations, tout particulièrement pour les filtres sur les variants. En effet quelle que soit la population, le nombre final de variants est à peu près équivalent (entre 535 000 et 545



(a) Les deux premières composantes principales calculées sur tout UK Biobank.



(b) Répartition des ethnies reportées.



(c) Les deux premières composantes principales fournies par UKBB pour les WB. Les individus dans l'ellipse d'équation $\frac{\left((x+12.5)\cos(-\frac{\pi}{8})+(y-3.3)\sin(-\frac{\pi}{8})\right)^2}{4.2^2} + \frac{\left((x+12.5)\sin(-\frac{\pi}{8})-(y-3.3)\cos(-\frac{\pi}{8})\right)^2}{3^2} = 1$ sont les VWB.

FIGURE 4.5 – Structure de population dans UK Biobank.

000). Pour les individus nous remarquons qu'il semble y avoir plus d'individus considérés comme apparentés parmi la population TTLM que pour les populations WB et VWB.

La différence la plus notable est les écarts-types plus élevés pour les approches à deux jeux de données que pour les approches avec projection pour les filtres sur les variants. Une explication est que l'ensemble de standardisation est plus petit que l'ensemble d'apprentissage, ce qui entraîne ces instabilités numériques.

4.5.3 Calcul des covariables de structure

Pour estimer la structure de population dans chaque échantillon avec l'ACP, une bonne pratique est d'utiliser des données élaguées [Anderson et al., 2010]. Ayant trouvé des différences dans la manière d'élaguer les données, nous avons décidé de comparer trois approches :

— Élagage des données en appliquant un "pruning" des variants avec un déséquilibre de liaison supérieur à 0.02 (au sens du r^2) et en supprimant manuellement deux régions à fort DL (la région 25-35 Mb sur le chromosome 6 et la région 54-60 Mb du chromosome 8) . Cela correspond aux QC trouvés dans Ge et al. [2017].

— Élagage des données en appliquant un "pruning" des variants avec une MAF < 0.05 et un déséquilibre de liaison supérieur à 0.05 au sens du r^2. On supprime également les régions à fort DL décrites dans la table 4.7 [Price et al., 2008, Bycroft et al., 2017]. Ces derniers correspondent plus au QC décrit dans la thèse de Claire Dandine-Roulland [Dandine-Roulland, 2014].

— Utilisation des composantes principales fournies par UK Biobank. Ces composantes principales ont été calculées avec $FastPCA$ [Galinsky et al., 2016] sur des données élaguées (plus de détails dans [Bycroft et al., 2017]).

Pour l'approche basée sur une projection, on calcule simplement la matrice de ressemblance sur $\mathcal{A}$ et pour l'approche avec deux jeux de données nous en calculerons pour les deux ensembles $\mathcal{A}$ et $\mathcal{S}$. Le nombre de variants intégrés dans chacune de ces trois approches est détaillé dans la table 4.6. Encore une fois les résultats sont très similaires entre les différentes approches pour intégrer les effets fixes, mais nous pouvons remarquer une plus grande instabilité numérique pour l'approche à deux jeux de données. Nous remarquons par contre que la deuxième méthode d'estimation des covariables inclut beaucoup moins de variants (environ moitié moins que les deux autres).

	TTLM		**WB**		**VWB**	
	Approche projection	Approche 2 jeux de données	Approche projection	Approche 2 jeux de données	Approche projection	Approche 2 jeux de données
Après échantionnage	$n_{\mathcal{A}} = 10\ 000$ - p = 784 256(0)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 784 256(0)	$n_{\mathcal{A}} = 10\ 000(0)$ - p = 784 256(0)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 784 256(0)	$n_{\mathcal{A}} = 10\ 000(0)$ - p = 784 256(0)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 784 256(0)
Callrate variants	$n_{\mathcal{A}} = 10\ 000(0)$ - p = 641 076(325)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 637 681(1 821)	$n_{\mathcal{A}} = 10\ 000(0)$ - p = 640 734(660)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 636 044(2 065)	$n_{\mathcal{A}} = 10\ 000(0)$ - p = 640 788(658)	$n_{\mathcal{A}} = 10\ 000(0)$ $n_{\mathcal{S}} = 1\ 000(0)$ p = 636 245(3 082)
Callrate individus	$n_{\mathcal{A}} = 9\ 696(11)$ - p = 641 076(325)	$n_{\mathcal{A}} = 9\ 696(14)$ $n_{\mathcal{S}} = 976(4)$ p = 637 681(1 821)	$n_{\mathcal{A}} = 9\ 677(21)$ - p = 640 734(660)	$n_{\mathcal{A}} = 9\ 680(20)$ $n_{\mathcal{S}} = 972(5)$ p = 636 044(2 065)	$n_{\mathcal{A}} = 9\ 686(18)$ - p = 640 788(658)	$n_{\mathcal{A}} = 9\ 690(19)$ $n_{\mathcal{S}} = 972(4)$ p = 636 245(3 082)
MAF	$n_{\mathcal{A}} = 9\ 696(11)$ - p = 549 927(303)	$n_{\mathcal{A}} = 9\ 696(14)$ $n_{\mathcal{S}} = 976(4)$ p = 545 983(1 698)	$n_{\mathcal{A}} = 9\ 677(21)$ - p = 549 6976(562)	$n_{\mathcal{A}} = 9\ 680(20)$ $n_{\mathcal{S}} = 972(5)$ p = 541 546(1 936)	$n_{\mathcal{A}} = 9\ 686(18)$ - p = 546 539(640)	$n_{\mathcal{A}} = 9\ 690(19)$ $n_{\mathcal{S}} = 972(4)$ p = 541 352(2 799)
Suppression non-bialléliques	$n_{\mathcal{A}} = 9\ 696(11)$ - p = 549 927(303)	$n_{\mathcal{A}} = 9\ 696(14)$ $n_{\mathcal{S}} = 976(4)$ p = 545 983(1 698)	$n_{\mathcal{A}} = 9\ 677(21)$ - p = 549 6976(562)	$n_{\mathcal{A}} = 9\ 680(20)$ $n_{\mathcal{S}} = 972(5)$ p = 541 546(1 936)	$n_{\mathcal{A}} = 9\ 686(18)$ - p = 546 539(640)	$n_{\mathcal{A}} = 9\ 690(19)$ $n_{\mathcal{S}} = 972(4)$ p = 541 352(2 799)
Test équilibre Hardy-Weinberg	$n_{\mathcal{A}} = 9\ 696(11)$ - p = 538 766(874)	$n_{\mathcal{A}} = 9\ 696(14)$ $n_{\mathcal{S}} = 976(4)$ p = 544 399(1 514)	$n_{\mathcal{A}} = 9\ 677(21)$ - p = 546 115(552)	$n_{\mathcal{A}} = 9\ 680(20)$ $n_{\mathcal{S}} = 972(5)$ p = 541 430(1 930)	$n_{\mathcal{A}} = 9\ 686(18)$ - p = 545 676(643)	$n_{\mathcal{A}} = 9\ 690(19)$ $n_{\mathcal{S}} = 972(4)$ p = 541 242(2 802)
Suppression individus apparentés	$n_{\mathcal{A}} = 8\ 929(19)$ - p = 538 766(874)	$n_{\mathcal{A}} = 8\ 889(19)$ $n_{\mathcal{S}} = 976(4)$ p = 544 399(1 514)	$n_{\mathcal{A}} = 9\ 396(27)$ - p = 546 115(552)	$n_{\mathcal{A}} = 9\ 359(20)$ $n_{\mathcal{S}} = 972(5)$ p = 541 430(1 930)	$n_{\mathcal{A}} = 9\ 448(21)$ - p = 545 676(643)	$n_{\mathcal{A}} = 9\ 409(17)$ $n_{\mathcal{S}} = 972(4)$ p = 541 242(2 802)

Tableau 4.2 – Effet des contrôles qualités sur les différentes populations et approches.

	TTLM		WB		VWB	
	Approche projection	Approche 2 études	Approche projection	Approche 2 études	Approche projection	Approche 2 études
Pruning r^2, 2 régions exclues	137 057(159)	136 675(119)	137 704(116)	135 271(292)	137 366(82)	135 114(281)
Pruning r^2, MAF, 23 régions exclues	75 442(55)	74 575(176)	75 756(80)	74 427(223)	75 632(49)	74 268(182)
CP fournies par UK Biobank	147 604(-)	147 604(-)	147 604(-)	147 604(-)	147 604(-)	147 604(-)

FIGURE 4.6 – Nombre de variants inclus dans les variables de structure pour chaque population.

Chromosome	Position (en Mb)
1	48-52
2	86-100.5 134.5-138 183-190
3	47.5-50 83.5-87 89-97.5
5	44-51.5 98-100.5 129-132 135.5-138.5
6	25-33.5 57-64 140-142.5
7	55-66
8	8-12 43-50
10	37-43
11	45-57 87.5-90.5
12	33-40 109.5-112
20	32-34.5

FIGURE 4.7 – Régions à fort DL

4.5.4 Prise en compte des covariables

La prise en compte des covariables non pénalisées pour l'approche avec une projection se fait avec une matrice de contraste et ne pose pas de difficulté particulière : la matrice de contraste permet de travailler sur un modèle sans covariables non-pénalisées et donc l'application de la régression ridge ne pose pas de difficultés. Pour l'approche avec un deuxième jeu de données, le processus est un petit peu plus compliqué et nous allons le décrire ici.

L'idée est de faire une régression linéaire du phénotype sur les covariables sur l'ensemble $\mathbf{S}$, de calculer les résidus du phénotype après soustraction des effets des covariables sur $\mathcal{A}$ en utilisant les coefficients calculés sur $\mathcal{S}$ puis d'utiliser la régression ridge sur ces résidus et les données génétiques. Il n'y a pas de difficulté particulière pour les covariables non pénalisées telles que le sexe ou l'âge mais pour les covariables de structure il y a un problème : les composantes principales de $\mathcal{S}$ ne sont pas comparables à celles de $\mathcal{A}$.

Pour "transférer" l'apprentissage de la structure de population de $\mathcal{S}$ à $\mathcal{A}$ nous utilisons les loadings calculés sur $\mathcal{S}$. Les loadings représentent l'influence des variants dans le calcul de composantes principales. Ils sont donc une quantité que l'on peut transférer d'un ensemble à un autre.

4.5.5 Estimations d'h^2

Nous allons ici regarder les estimations d'h^2 pour les différents phénotypes, populations, méthodologies d'estimation d'héritabilité et d'intégration des covariables. Les résultats sont présentés pour respectivement la taille, l'IMC, la circonférence et le tour de taille dans les figures 4.8, 4.9, 4.10 et 4.11.

Un premier résultat est que les estimations moyennes sont plutôt en accord avec la littérature dans tous les cas [Ge et al., 2017], ce qui est rassurant. Nous allons maintenant décrire les différences et les sources de variance entre les différentes estimations.

Pour commencer nous observons sans grande surprise des différences d'estimation d'héritabilité selon la population choisie. Ces différences ne semblent dépasser plus de 5 points de pourcentage et restent donc assez minimes. Ce résultat semble dire que l'héritabilité des ces traits est relativement constante entre différentes populations. Il convient toutefois de rester prudent : les individus déclarés comme caucasiens repré-

sentent plus de 90 % de UKBB et donc même la population TTLM est très majoritairement composée de personnes caucasiennes. Cela pourrait donc expliquer les fortes ressemblances d'estimation entre les différentes populations.

Le choix de la méthodologie pour estimer l'héritabilité (régression ridge avec GCV ou AIREML) ne semble pas non plus être une grande source de variabilité. A l'exception de la taille où l'on observe des différences allant jusqu'à 5 points de pourcentage le choix de la méthodologie ne semble pas donner des résultats très différents. C'est un résultat rassurant qui confirme nos résultats sur les simulations. Nous remarquons que pour la taille (i.e. une héritabilité forte) la GCV renvoie des résultats moins élevés que l'AI-REML et que c'est l'inverse pour les autres phénotypes (dont l'héritabilité est plus faible), mais ce phénomène est très minime. Signalons également qu'avec seulement 10 échantillons il sera difficile d'avoir une conclusion arrêtée.

Nous avons observé une influence significative du choix de l'approche pour intégrer les covariables (avec une matrice de projection ou un deuxième jeu de données) sur chacune des deux méthodes d'estimations d'héritabilité (AI-REML et GCV). Les estimations d'héritabilité semblent plus élevées quand nous utilisons l'approche à deux jeux de données (sauf pour la taille où c'est l'inverse), en particulier quand on utilise les CP de UKBB. Le cas de la circonférence des hanches pour la population TTLM et utilisant les CP de UKBB est l'exemple le plus marquant de différence d'estimation d'héritabilité entre les deux approches. Une explication pourrait être trouvée en enquêtant sur l'association des variants et des covariables.

La manière d'approcher les covariables de structure semble néanmoins être source de peu de variabilité : à population, méthode d'estimation et approche pour intégrer les covariables fixées, les estimations d'héritabilité semblent peu affectées par la stratégie de construction des CP. C'est un résultat qui nous rassure sur l'impression d'absence de consensus dans la construction de ces covariables puisque les estimations restent cohérentes. Notons toutefois une exception pour l'estimation d'héritabilité de la circonférence des hanches : sur la population TTLM et en utilisant l'approche à deux études pour intégrer les covariables, les estimations d'héritabilité (réalisées avec l'AI-REML ou la GCV) sont bien plus élevées si nous utilisons les PC de UKBB plutôt que les deux autres.

La plus grande source de variabilité de nos estimations est l'échantillonnage des données. Par exemple pour la taille nous observons qu'en fixant la population, la méthode

pour estimer l'héritabilité, l'approche pour intégrer les covariables et la stratégie de construction des CP, nous pouvons observer des différences entre estimations d'héritabilité de plus de 10 points de pourcentage. Ce résultat est clairement le moins rassurant pour l'héritabilité et interroge sur les résultats trouvés dans la littérature.

4.6 En résumé ...

Nous avons dans ce chapitre présenté le problème d'estimation de l'héritabilité comme un problème d'apprentissage. Nous avons proposé deux quantités dérivées de la prédiction pour estimer l'héritabilité : une première nommée h_r^2 basée sur l'estimation de la variance du vecteur de prédiction du terme génétique et une seconde nommée h_p^2 (et ici confondue avec le coefficient de détermination R^2) définie comme une quantité mesurant la qualité de la prédiction.

Nous avons regardé les capacités d'h_r^2 et d'h_p^2 pour estimer l'héritabilité à l'aide de simulations. Les deux quantités utilisaient des prédictions réalisées avec la régression ridge. Ces quantités ont été calculées avec et sans un ensemble de test. Nous avons également proposé une méthode d'estimation d'héritabilité basée sur une transformation du paramètre de pénalisation de la régression ridge pour obtenir h_g^2.

Les résultats sur simulations ont montré qu'h_r^2 et h_p^2 ne donnaient pas d'estimation d'héritabilité satisfaisante en grande dimension : avec un ensemble de test les prédictions étaient de très mauvaise qualité et sans cet ensemble de test il y avait un fort surapprentissage qui biaisait les quantités. En revanche l'estimation d'h_g^2 en passant par la transformation du paramètre de pénalisation de la régression ridge a donné de bons résultats, en accord avec les méthodes de référence.

Nous avons donc essayé notre méthode d'estimation d'h_g^2 sur les données de UK-Biobank. Ici encore, les estimations d'héritabilité étaient plutôt satisfaisantes et en accord avec la méthode de référence.

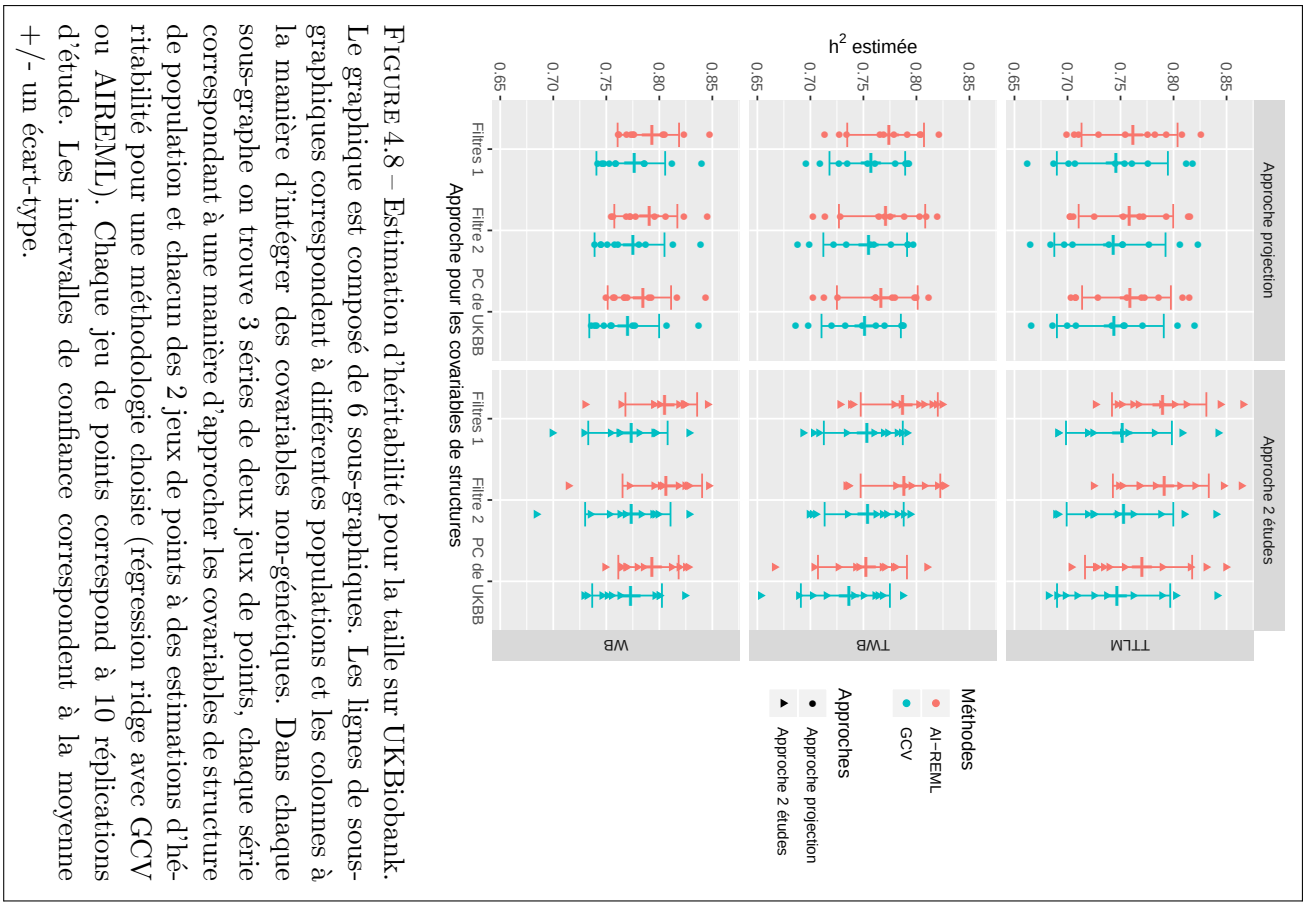


FIGURE 4.8 – Estimation d'héritabilité pour la taille sur UKBiobank. Le graphique est composé de 6 sous-graphiques. Les lignes de sous-graphiques correspondent à différentes populations et les colonnes à la manière d'intégrer des covariables non-génétiques. Dans chaque sous-graphe on trouve 3 séries de deux jeux de points, chaque série correspondant à une manière d'approcher les covariables de structure de population et chacun des 2 jeux de points à des estimations d'héritabilité pour une méthodologie choisie (régression ridge avec GCV ou AIREML). Chaque jeu de points correspond à 10 réplications d'étude. Les intervalles de confiance correspondent à la moyenne +/- un écart-type.

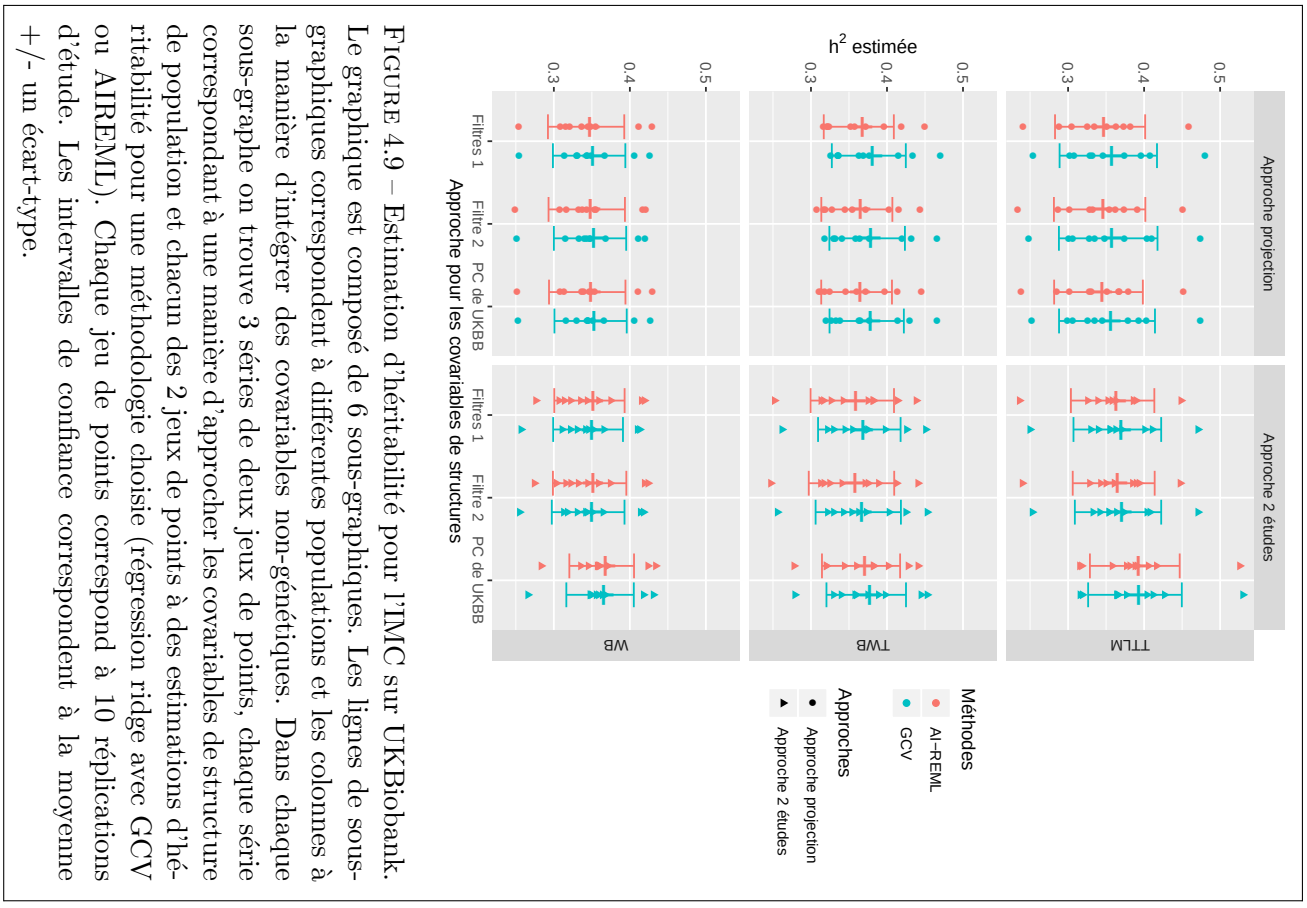


FIGURE 4.9 – Estimation d'héritabilité pour l'IMC sur UKBiobank. Le graphique est composé de 6 sous-graphiques. Les lignes de sous-graphiques correspondent à différentes populations et les colonnes à la manière d'intégrer des covariables non-génétiques. Dans chaque sous-graphe on trouve 3 séries de deux jeux de points, chaque série correspondant à une manière d'approcher les covariables de structure de population et chacun des 2 jeux de points à des estimations d'héritabilité pour une méthodologie choisie (régression ridge avec GCV ou AIREML). Chaque jeu de points correspond à 10 réplications d'étude. Les intervalles de confiance correspondent à la moyenne +/- un écart-type.

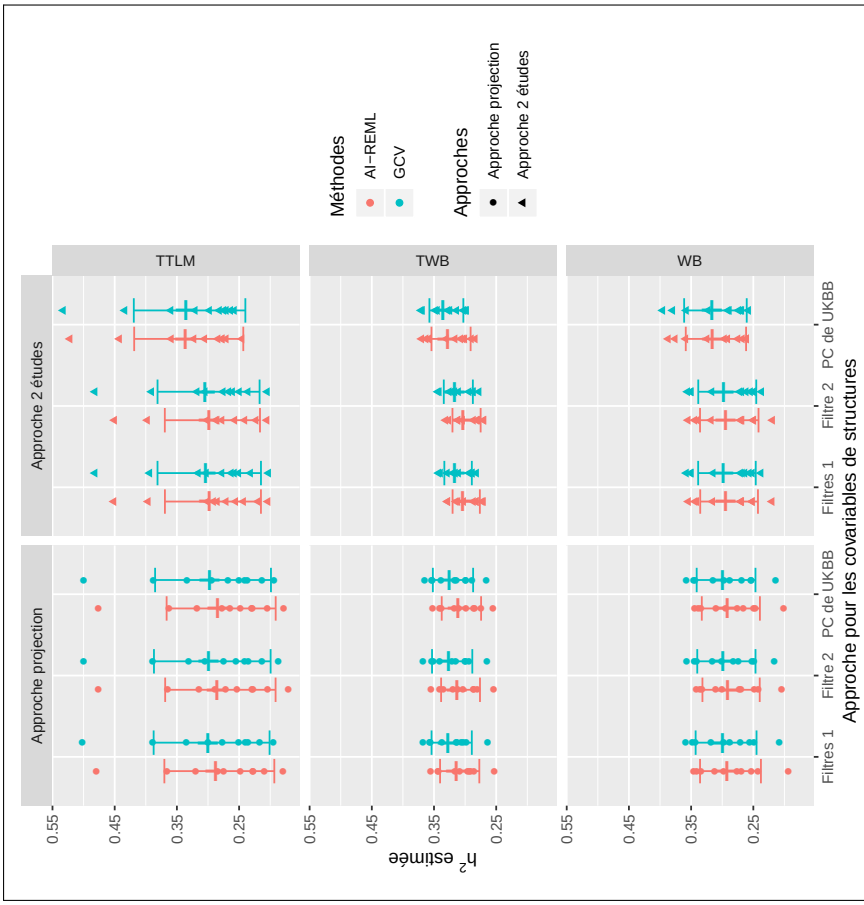


FIGURE 4.11 – Estimation d'héritabilité pour le tour de taille sur UKBiobank. Le graphique est composé de 6 sous-graphiques. Les lignes de sous-graphiques correspondent à différentes populations et les colonnes à la manière d'intégrer des covariables non-génétiques. Dans chaque sous-graphe on trouve 3 séries de deux jeux de points, chaque série correspondant à une manière d'approcher les covariables de structure de population et chacun des 2 jeux de points à des estimations d'héritabilité pour une méthodologie choisie (régression ridge avec GCV ou AIREML). Chaque jeu de points correspond à 10 réplications d'étude. Les intervalles de confiance correspondent à la moyenne +/- un écart-type.

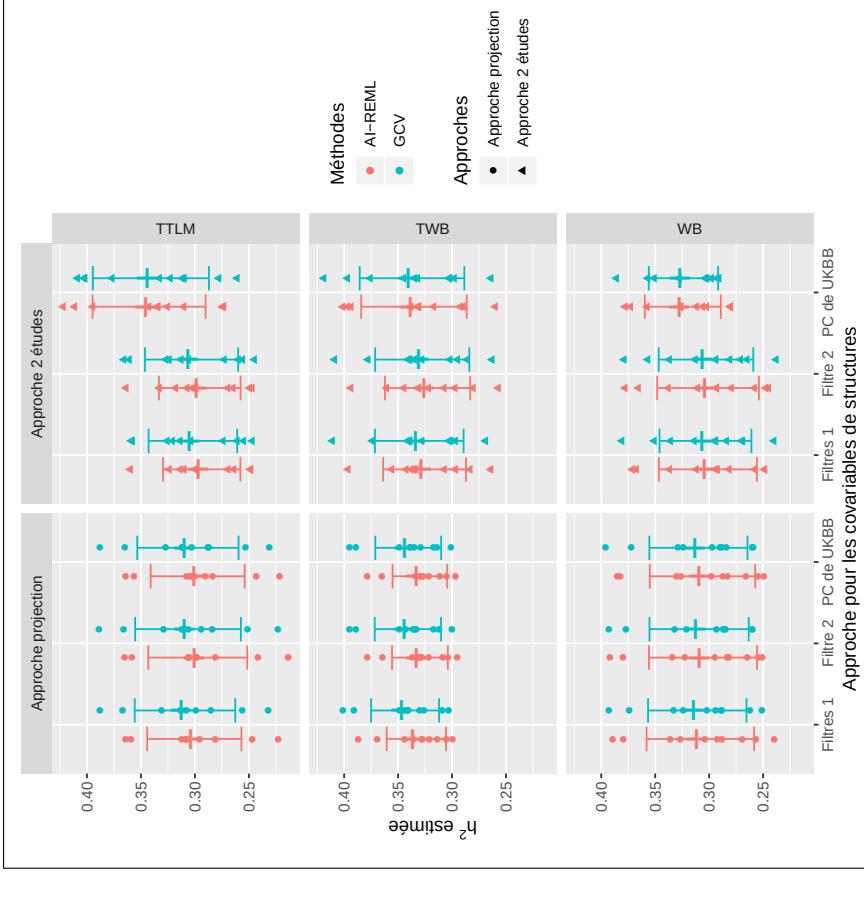


FIGURE 4.10 – Estimation d'héritabilité pour la circonférence des hanches sur UKBiobank. Le graphique est composé de 6 sous-graphiques. Les lignes de sous-graphiques correspondent à différentes populations et les colonnes à la manière d'intégrer des covariables non-génétiques. Dans chaque sous-graphe on trouve 3 séries de deux jeux de points, chaque série correspondant à une manière d'approcher les covariables de structure de population et chacun des 2 jeux de points à des estimations d'héritabilité pour une méthodologie choisie (régression ridge avec GCV ou AIREML). Chaque jeu de points correspond à 10 réplications d'étude. Les intervalles de confiance correspondent à la moyenne +/- un écart-type.

Chapitre 5

Pouvoir prédictif de la régression ridge

Dans ce chapitre nous nous intéressons au comportement de l'erreur de prédiction de l'estimateur ridge selon les dimensions de notre problème. Nous commencerons par proposer un bref aperçu de la littérature, puis nous présenterons l'idée de notre approximation et son application sur le MSE et le carré de la corrélation puis enfin nous vérifierons la validité de notre approximation sur des simulations et des données réelles.

5.1 Contexte

De nombreux auteurs se sont penchés sur la précision des prédictions en sélection génomique. La littérature définit cette précision comme le carré de la corrélation entre le phénotype et sa prédiction, ce qui n'est pas la définition la plus classique en apprentissage statistique (où on préférera l'erreur quadratique moyenne de prédiction). Plusieurs auteurs se sont penchés sur l'écriture de formules permettant d'approcher cette précision en la liant à des quantités telles que la taille de l'ensemble d'apprentissage, le nombre de variants ou encore l'héritabilité [Brard and Ricard, 2015]. Daetwyler et al. [2008] donnent une formule de la corrélation entre les effets génétiques réels et les effets génétiques estimés avec des régressions linéaires univariées sous hypothèse d'effets fixes, supposant l'indépendance des variants causaux. Ils utilisent des régressions linéaires univariées pour chacun des variants puis les combinent après, ce qui est équivalent au calcul d'un *Polygenic Risk Score* (PRS) (voir Pharoah et al. [2002], Purcell et al. [2009]). Les auteurs proposent d'approximer le pouvoir prédictif par $\left(h^2/(h^2+\frac{p}{n})\right)^2$ avec n la taille de l'échantillon d'apprentissage $\mathcal{A}$, p le nombre de variants indépendants et h^2 l'héritabilité pour des phénotypes quantitatifs et qualitatifs. Notons qu'à taille de $\mathcal{A}$

fixée et à cause du DL, l'intégration d'un grand nombre de loci dans l'étude compromet l'hypothèse d'indépendance et que cette formule semble mal se prêter au cadre des GWAS. Goddard [2009] étend cette formule aux GBLUP en remplaçant le concept de nombre de variants indépendants par le nombre effectif de variants indépendants (aussi appelé le nombre effectif de segments de chromosome indépendants et noté M_e). Les auteurs proposent d'approximer le pouvoir prédictif par $\sqrt{1-\frac{\lambda}{2n\sqrt{a}}\log(\frac{1+a+2\sqrt{a}}{1+a-2\sqrt{a}})}$ avec $\lambda = \frac{M_e}{h^2\log(2N_e)}$ et $a = 1+2\frac{M_e}{nh^2\log(2N_e)}$, où N_e représente la taille de population effective. Par la suite une très riche littérature s'est développée autour de ces formules et pour l'estimation de M_e (voir [Brard and Ricard, 2015] pour une synthèse). Daetwyler et al. [2010] actualisent la formule proposée dans [Daetwyler et al., 2008] en remplaçant le nombre total de loci par M_e. Rabier et al. [2016] proposent, en supposant que l'on dispose de variants en DL avec les variants causaux, une approximation du R^2 par $\sqrt{h^2}\sqrt{\frac{h^2/(1-h^2)}{\mathbb{E}\left[\left\|z_{te}^T\mathbf{Z}^T\mathbf{V}^{-1}\right\|^2\right]+\frac{h^2}{1-h^2}}}$, avec z_{te} représentant un individu de test indépendant de l'ensemble d'apprentissage, $\mathbf{Z}$ la matrice de marqueurs correspondant à l'ensemble d'apprentissage (que l'on considère comme non-aléatoire) et $\mathbf{V} = \mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I}_n$. Elsen [2017] utilise un développement de Taylor (d'ordre 1) et propose comme approximation $\frac{nh^2}{nh^2+p(1-h^2)}$ dépendant uniquement de l'héritabilité, du nombre de variants génotypés et de la taille de $\mathcal{A}$ dans un cadre de petite dimension. Notons que l'auteur propose également une formule (plus complexe) avec un développement de Taylor du second ordre.

de Vlaming and Groenen [2015] ont montré à l'aide de simulations un fort lien entre PRS et régression ridge dans le cadre des GWAS (i.e. dans un cadre où l'ensemble d'apprentissage est de "petite" taille). Les auteurs ont également montré que les capacités prédictives de la régression ridge s'améliorent quand la taille de l'ensemble d'apprentissage augmente et deviennent meilleures que celles des PRS.

De los Campos et al. [2013] se sont intéressés au pouvoir prédictif des GBLUP pour la génétique humaine. Comme pour [Rabier et al., 2016], les auteurs ont utilisé un modèle avec les variants causaux et un modèle avec les marqueurs génotypés. Notons enfin que dans leur article les auteurs ont montré que le carré de la corrélation entre le phénotype et le GBLUP tend vers l'héritabilité si les individus sont non-apparentés. Dans leur article les auteurs restent toutefois très pessimistes sur la capacité des GBLUP à effectuer de bonnes prédictions sur des données humaines de génotypage du fait du DL incomplet entre marqueurs génotypés et variants causaux. Ils ont en particulier montré qu'en absence de fort DL entre les variants causaux et génotypés, les données

de pédigrés amènent de meilleures capacités prédictives que les données de population.

Dandine-Roulland and Perdry [2015] proposent d'approximer le coefficient de détermination R^2 (calculé sur un ensemble de test) de l'estimateur des BLUP sur des données humaines par $n(h^2)^2\nu$, une formule linéaire en la taille de $\mathcal{A}$ et quadratique en l'héritabilité mais également fonction de ν la variance des coefficients de la matrice $\mathbf{Z}\mathbf{Z}_{te}^T$ (avec $\mathbf{Z}_{te}^T$ la matrice de génotype d'un échantillon de test) qui est sûrement une fonction de p. Les auteurs restent également très pessimistes sur la capacité des BLUP à prédire efficacement un phénotype complexe chez l'humain.

Zhao and Zhu [2019] proposent des formules du R^2 (non détaillées ici) pour la prédiction croisée de phénotypes. Les auteurs se sont intéressés aux erreurs sur des ensembles de test mais également d'apprentissage, et ils ont lié l'estimateur marginal (le PRS) et l'estimateur de la régression ridge.

Nous allons ici présenter une approximation simple de l'erreur de prédiction pour l'estimateur de la régression ridge. Notre approximation n'a pas pour objectif d'approcher la vérité pour le vrai modèle biologique sous-jacent du phénotype mais plutôt de donner une bonne intuition au lecteur de pourquoi le pouvoir prédictif de la régression ridge (ou les GBLUP plus généralement) est souvent décevant en génétique humaine pour des individus non-apparentés.

5.2 Une approximation de pouvoir prédictif selon le rapport $\frac{n}{p}$ pour des données de GWAS

5.2.1 Idée de l'approximation

Nous proposons dans cette section des approximations afin d'écrire l'erreur comme une fonction du ratio $\frac{n}{p}$. Nous décrivons rapidement les approximations que nous allons faire ici.

Supposons que les individus n'ont pas de lien de parenté donc la matrice de covariance des individus est diagonale. La matrice de covariance des variants est également diagonale puisque l'on fait l'hypothèse de l'indépendance des variants. Supposons également que les données sont normalisées, alors les matrices $\mathbf{Z}\mathbf{Z}^T$ et $\mathbf{Z}^T\mathbf{Z}$ sont les matrices de covariance empirique des individus et des variants respectivement (à un facteur multiplicatif p ou n près). Nous proposons donc une approximation selon la valeur du ratio n/p :

— Pour la grande dimension ($p > n$) la matrice $\mathbf{Z}\mathbf{Z}^T$ estime bien la matrice de covariance des individus (à un facteur p près).

— A l'inverse $\mathbf{Z}^T\mathbf{Z}$ est un bon estimateur de la matrice de covariance des variants quand $n > p$.

En résumé, nous utiliserons l'approximation $\mathbf{Z}\mathbf{Z}^T \simeq p\mathbf{I}_n$ quand $n < p$ et $\mathbf{Z}^T\mathbf{Z} \simeq n\mathbf{I}_p$ quand $n > p$.

Posons également les hypothèses suivantes :

— $\forall i \in [\![1, n]\!]$ $\text{var}(y_i) = 1$, alors on a $\sigma^2 = 1 - h^2$.

— l'héritabilité est "équitablement répartie entre les variants" i.e. $\forall j \in [\![1, p]\!]$ $\text{var}(u_j) = \frac{h^2}{p}$ (ce qui correspond aux hypothèses du modèle mixte). Nous pourrons raisonnablement approximer l'héritabilité par l'héritabilité génomique et donc utiliser les liens entre λ et h^2.

— $u^T u \simeq p \times \frac{h^2}{p}$ et $(\mathbf{Z}u)^T(\mathbf{Z}u) \simeq nh^2$.

Nous allons appliquer notre approximation sur trois quantités : l'erreur de prédiction calculée pour un individu de test, l'erreur de prédiction calculée sur l'ensemble d'apprentissage et enfin le carré de la corrélation calculé pour un individu de test. Nous détaillerons les calculs pour l'erreur de prédiction pour un individu de test uniquement, et discuterons directement de l'approximation pour les autres quantités. Les détails des ces dernières seront en annexe.

5.2.2 Pour l'erreur de prédiction sur l'ensemble de test

Écriture du dilemme biais-variance

Nous supposerons que notre phénotype suit les hypothèses du modèle polygénique additif avec tous les variants causaux pour un jeu d'individus non-apparentés

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e}. \tag{5.1}$$

avec $\mathbf{Z}$ centrée et réduite par colonnes et $\mathbf{e} \sim \mathcal{N}(0_n, \sigma^2\mathbf{I}_n)$. On rappelle que $\hat{u} = \mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I}_n\right)^{-1}\mathbf{y}_{tr} = \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}^T\mathbf{y}_{tr} = \mathbf{K}_\lambda\mathbf{y}_{tr}$.

Nous allons regarder le comportement de l'erreur de prédiction de l'estimateur ridge sans fixer les données de test. Pour la suite des calculs, l'indice $_{tr}$ fera référence à

l'ensemble d'apprentissage et l'indice $_{te}$ à un individu de test. En supposant que la matrice de génotypes de l'ensemble d'apprentissage est fixée, la formule d'espérance totale nous permet d'écrire

$$\begin{aligned}\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] &= \mathbb{E}_{z_{te}}\left[\mathbb{E}_{\mathbf{y}_{tr},y_{te}|z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right]\right] \\ &= \mathbb{E}_{z_{te}}\left[\mathrm{var}(y_{te}|z_{te})+\mathrm{var}(\hat{y}_{te}|z_{te})+\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]-\mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right].\end{aligned}$$

Nous montrons facilement que $\mathbb{E}_{y_{te}|z_{te}}\left[y_{te}\right]=z_{te}^T u$ et $\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]=z_{te}^T\mathbf{K}_\lambda\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\mathbf{y}_{tr}\right]=z_{te}^T\mathbf{K}_\lambda(\mathbf{Z}u)$. Nous pouvons alors écrire les termes de la décomposition biais-variance sous les formes suivantes :

$$\begin{aligned}\mathrm{var}(y_{te}|z_{te}) &= \mathbb{E}_{y_{te}|z_{te}}\left[\left(y_{te}-\mathbb{E}_{y_{te}|z_{te}}\left[y_{te}\right]\right)^2\right]=\mathbb{E}_{y_{te}|z_{te}}\left[e_{te}^2\right]=\sigma^2 \\ \mathrm{var}(\hat{y}_{te}|z_{te}) &= \mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}^2\right]-\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]^2 \\ &= \mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\mathbf{y}_{tr}^T\mathbf{K}_\lambda^T z_{te}z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}\right]-(\mathbf{Z}u)^T\mathbf{K}_\lambda^T z_{te}z_{te}^T\mathbf{K}_\lambda(\mathbf{Z}u) \\ &= \mathrm{tr}\left(\mathbf{K}_\lambda^T z_{te}z_{te}^T\mathbf{K}_\lambda\sigma^2\mathbf{I}_n\right)+(\mathbf{Z}u)^T\mathbf{K}_\lambda^T z_{te}z_{te}^T\mathbf{K}_\lambda(\mathbf{Z}u) \\ &\quad -(\mathbf{Z}u)^T\mathbf{K}_\lambda^T z_{te}z_{te}^T\mathbf{K}_\lambda(\mathbf{Z}u) \\ &= \sigma^2 z_{te}^T\mathbf{K}_\lambda\mathbf{K}_\lambda^T z_{te} \\ \left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]-\mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2 &= \left(z_{te}^T\mathbf{K}_\lambda(\mathbf{Z}u)-z_{te}^T u\right)^2 \\ &= \left(z_{te}^T\left(\mathbf{K}_\lambda\mathbf{Z}-\mathbf{I}_p\right)u\right)^2 \\ &= z_{te}^T\left(\mathbf{K}_\lambda\mathbf{Z}-\mathbf{I}_p\right)uu^T\left(\mathbf{K}_\lambda\mathbf{Z}-\mathbf{I}_p\right)z_{te}.\end{aligned}$$

En appliquant l'hypothèse que tous les variants sont indépendants et de même variance $\mathbb{E}_{z_{te}}\left[z_{te}\right]=0_p$ et $\mathrm{var}(z_{te})=\mathbf{I}_p$, nous obtenons

$$\begin{aligned}\mathbb{E}_{z_{te}}\left[\mathrm{var}(y_{te}|z_{te})\right] &= \sigma^2 \\ \mathbb{E}_{z_{te}}\left[\mathrm{var}(\hat{y}_{te}|z_{te})\right] &= \sigma^2\mathrm{tr}\left(\mathbf{K}_\lambda\mathbf{K}_\lambda^T\right) \\ \mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]-\mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right] &= u^T\left(\mathbf{K}_\lambda\mathbf{Z}-\mathbf{I}_p\right)^2 u \\ &= u^T\left(\mathbf{K}_\lambda\mathbf{Z}\mathbf{K}_\lambda\mathbf{Z}-2\mathbf{K}_\lambda\mathbf{Z}+\mathbf{I}_p\right)u.\end{aligned}$$

Notons que le premier terme n'est PAS la variance totale puisque qu'il y a un conditionnement par z_{te}. Les deux autres termes seront dans la suite abusivement

appelés variance et biais.

Le cas $p > n$

Dans le cadre $n < p$ nos approximations nous permettent d'écrire

$$\mathbf{Z}\mathbf{Z}^T \simeq p\mathbf{I}_n \Rightarrow \mathbf{K}_\lambda \simeq \frac{1}{p+\lambda}\mathbf{Z}^T.$$

Cette approximation revient à supposer que la matrice $\mathbf{Z}\mathbf{Z}^T$ possède n valeurs propres égales et valant p. Nous pouvons alors écrire les termes de l'erreur selon le ratio $\frac{n}{p}$.

$$\begin{aligned}
\mathbb{E}_{z_{te}}\left[\mathrm{var}(\hat{y}_{te}|z_{te})\right] &= \sigma^2 \mathrm{tr}\left(\mathbf{K}_\lambda \mathbf{K}_\lambda^T\right) \\
&\simeq \sigma^2 \mathrm{tr}\left(\left(\frac{1}{p+\lambda}\mathbf{Z}^T\right)\left(\frac{1}{p+\lambda}\mathbf{Z}^T\right)^T\right) \\
&= \sigma^2 \left(\frac{1}{p+\lambda}\right)^2 \mathrm{tr}\left(\mathbf{Z}^T\mathbf{Z}\right) \\
&\simeq \sigma^2 \left(\frac{1}{p+\lambda}\right)^2 \mathrm{tr}\left(p\mathbf{I}_n\right) \\
&= \sigma^2 \left(\frac{1}{p+\lambda}\right)^2 np \\
&= (1-h^2)(h^2)^2\frac{n}{p}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right] - \mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right] &= u^T\left(\mathbf{K}_\lambda\mathbf{Z} - \mathbf{I}_p\right)^2 u \\
&= u^T\left(\left(\frac{1}{p+\lambda}\right)^2 \mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T\mathbf{Z} - 2\left(\frac{1}{p+\lambda}\right)\mathbf{Z}^T\mathbf{Z} + \mathbf{I}_p\right)u \\
&= p\left(\frac{1}{p+\lambda}\right)^2 (\mathbf{Z}u)^T(\mathbf{Z}u) - 2\left(\frac{1}{p+\lambda}\right)(\mathbf{Z}u)^T(\mathbf{Z}u) + u^Tu \\
&\simeq p\left(\frac{1}{p+\lambda}\right)^2 nh^2 - 2\left(\frac{1}{p+\lambda}\right)nh^2 + h^2 \\
&= \frac{n}{p}(h^2)^3 - 2\frac{n}{p}(h^2)^2 + h^2 \\
&= h^2\left(1 + \frac{n}{p}\left((h^2)^2 - 2h^2\right)\right).
\end{aligned}$$

En additionnant les termes d'erreur irréductible, de biais et de variance, nous obtenons l'approximation suivante de l'erreur quadratique :

$$\begin{aligned}\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] &\simeq 1-h^2+(1-h^2)(h^2)^2\frac{n}{p}+h^2\left(1+\frac{n}{p}\left((h^2)^2-2h^2\right)\right)\\ &= 1+h^2\left(-1+(1-h^2)h^2\frac{n}{p}+1+\frac{n}{p}\left((h^2)^2-2h^2\right)\right)\\ &= 1-\frac{n}{p}(h^2)^2.\end{aligned}$$

Le cas $n > p$

Dans le cadre $n > p$ nos approximations nous permettent d'écrire

$$\mathbf{Z}^T\mathbf{Z} \simeq n\mathbf{I}_p \Rightarrow \mathbf{K}_\lambda \simeq \frac{1}{n+\lambda}\mathbf{Z}^T$$

Avec un peu d'algèbre nous avons les égalités suivantes (voir leurs démonstrations en annexe C.1.1) :

$$\frac{n}{n+\lambda} = \frac{\frac{n}{p}\times h^2}{1+h^2\times(\frac{n}{p}-1)} \tag{5.2}$$

$$\frac{\lambda}{n+\lambda} = \frac{1-h^2}{1+h^2(\frac{n}{p}-1)} \tag{5.3}$$

Nous pouvons alors approximer les termes de biais et de variance :

$$\begin{aligned}\mathbb{E}_{z_{te}}\left[\mathrm{var}(\hat{y}_{te}|z_{te})\right] &= \sigma^2\mathrm{tr}\left(\mathbf{K}_\lambda\mathbf{K}_\lambda^T\right)\\ &\simeq \sigma^2\mathrm{tr}\left(\left(\frac{1}{n+\lambda}\right)^2\mathbf{Z}^T\mathbf{Z}\right)\\ &\simeq \sigma^2\left(\frac{1}{n+\lambda}\right)^2 np\\ &= \sigma^2\frac{1}{\frac{n}{p}}\left(\frac{\frac{n}{p}\times h^2}{1+h^2\times(\frac{n}{p}-1)}\right)^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]-\mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right] &= u^T\left(\mathbf{K}_\lambda\mathbf{Z}\mathbf{K}_\lambda\mathbf{Z}-2\mathbf{K}_\lambda\mathbf{Z}+\mathbf{I}_p\right)u\\ &\simeq u^T\left(\left(\frac{1}{n+\lambda}\right)^2\mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T\mathbf{Z}-2\left(\frac{1}{n+\lambda}\right)\mathbf{Z}^T\mathbf{Z}+\mathbf{I}_p\right)u\\ &\simeq \left(\frac{n}{n+\lambda}-1\right)^2 u^T u \simeq \left(\frac{n}{n+\lambda}-1\right)^2 h^2\\ &= \left(\frac{1-h^2}{1+h^2(\frac{n}{p}-1)}\right)^2 h^2\end{aligned}$$

En sommant ces expressions et en simplifiant, nous arrivons à l'expression de l'erreur quadratique suivante :

$$\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] \simeq (1-h^2)\frac{1+\frac{n}{p}h^2}{1+h^2(\frac{n}{p}-1)}$$

En résumé, après avoir groupé les expressions de l'erreur quadratique, de la variance et du carré du biais, nous avons

$$\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] \simeq \begin{cases} 1-\frac{n}{p}(h^2)^2 \text{ si } n<p \\ (1-h^2)\frac{1+\frac{n}{p}h^2}{1+h^2(\frac{n}{p}-1)} \text{ sinon.}\end{cases} \tag{5.4}$$

$$\mathbb{E}_{z_{te}}\left[\mathrm{var}(\hat{y}_{te}|z_{te})\right] \simeq \begin{cases} (1-h^2)(h^2)^2\frac{n}{p} \text{ si } n<p \\ (1-h^2)\frac{1}{\frac{n}{p}}\left(\frac{\frac{n}{p}\times h^2}{1+h^2\times(\frac{n}{p}-1)}\right)^2 \text{ sinon.}\end{cases} \tag{5.5}$$

$$\mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right]-\mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right] \simeq \begin{cases} h^2\left(1+\frac{n}{p}\left((h^2)^2-2h^2\right)\right) \text{ si } n<p \\ \left(\frac{1-h^2}{1+h^2(\frac{n}{p}-1)}\right)^2 h^2 \text{ sinon.}\end{cases} \tag{5.6}$$

Nous avons vérifié en annexe C.1.2 que les limites de ces quantités sont cohérentes.

Nous avons tracé en 5.1 les courbes des trois quantités que nous avons approchées selon différentes valeurs d'héritabilité. Commençons par voir que l'erreur (figure 5.1a) est une fonction décroissante en n/p et qui tend vers l'erreur irréductible quand $\frac{n}{p} \to +\infty$, ce qui est un comportement normal d'un point de vue statistique : plus on a d'individus et plus l'erreur de prédiction est censée diminuer jusqu'à atteindre son erreur minimale qu'est l'erreur irréductible. L'autre cas extrême donne un résultat plus surprenant. En effet nous voyons que même pour une héritabilité de 1, l'erreur tend vers 1 quand $\frac{n}{p} \to 0$. Autrement dit même dans un cadre où il n'y a aucun aléa, il

est impossible de faire de bonne prédiction quand $n << p$. Une interprétation est que quand le ratio n/p devient trop faible, l'estimateur de la ridge tend vers un vecteur nul. Notons également que l'approximation est croissante en h^2.

Nous allons maintenant décrire le carré du biais (figure 5.1b) et la variance (figure 5.1c). Nous voyons que le carré du biais est une fonction décroissante en n/p, qui tend vers h^2 quand $n/p \to 0$ et vers 0 quand $n/p \to +\infty$.

La variance est une fonction non monotone de n/p qui a un maximum dans $]1, +\infty[$ si $h^2 < 2/3$ (cette valeur est obtenue par la dérivation de l'approximation de la variance quand $n/p > 1$ que nous ne détaillerons pas ici) tandis que le maximum est en $n/p = 1$ si $h^2 \geq 2/3$. Elle tend à droite et à gauche vers 0.

Les limites à droite du carré du biais et de la variance ne sont pas surprenantes : plus on a d'individus, moins la régression ridge a besoin de s'éloigner de l'estimateur des moindres carrés (qui est sans biais). Il est donc cohérent que le carré du biais et la variance tendent vers 0 quand $n/p \to +\infty$.

Lorsque $p >> n$, pour une valeur de λ optimale (proportionnelle à p), le biais domine la variance. L'estimateur des paramètres de la régression ridge tend vers le vecteur nul et sa variance vers 0. Ainsi, de par l'équation du compromis biais-variance, le carré du biais tend vers l'héritabilité car la variance totale du phénotype (et donc l'erreur dans notre cas) vaut l'unité.

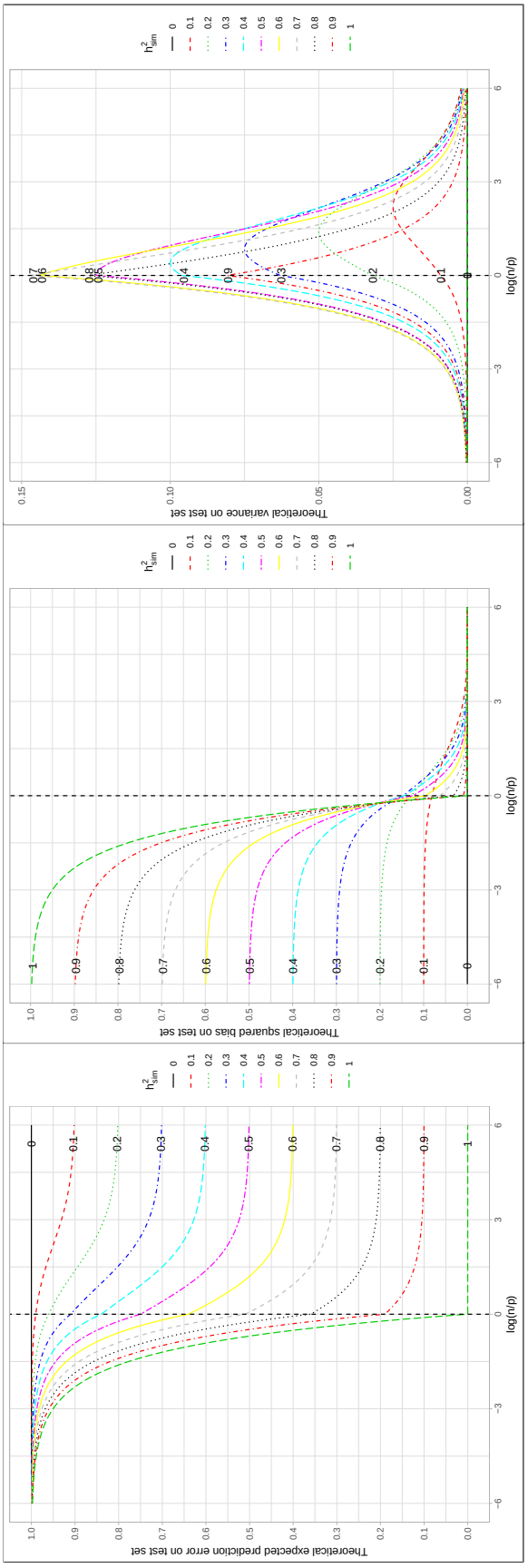
Dans la section 5.3, nous proposerons une interprétation de cette approximation de la régression ridge pour la voir comme une combinaison d'estimateurs des marginales et reviendrons sur ces résultats.

5.2.3 Pour l'erreur de prédiction sur l'ensemble d'apprentissage

En utilisant notre approximation, il est possible d'écrire l'erreur de prédiction sur l'ensemble d'apprentissage comme :

$$\mathbb{E}_{\mathbf{y}_{tr}}\left[\frac{1}{n}(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})^T(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})\right] \simeq \begin{cases} (1-h^2)^2 \text{ si } n < p \\ 1 - 2\frac{n}{n+\lambda}\left(\frac{p}{n}(1-h^2)+h^2\right) + \left(\frac{n}{n+\lambda}\right)^2\left(\frac{p}{n}(1-h^2)+h^2\right) \text{ sinon.} \end{cases}$$

Comme pour l'erreur sur le test, nous avons tracé un graphe de l'erreur sur l'ensemble d'apprentissage pour différentes valeurs d'h^2 dans la figure 5.2. Un résultat très surprenant est que l'erreur ne semble pas dépendre du ratio n/p quand celui-ci est



(a) Erreur (b) Carré du biais (c) Variance

FIGURE 5.1 – Erreur quadratique, carré du biais et variance théoriques avec notre approximation sur l'ensemble de test selon le logarithme du ratio n/p en supposant que la variance totale vaut 1. Chaque courbe correspond à une héritabilité simulée.

inférieur à 1. A l'inverse quand ce dernier est supérieur à 1, l'erreur est une fonction croissante qui tend vers l'erreur irréductible. Ce comportement démontre que le surapprentissage est visible dans l'erreur calculée sans ensemble de test puisque l'on arrive à faire "mieux" que l'erreur irréductible. Remarquons également que la différence de l'erreur théorique entre l'ensemble de test et l'ensemble d'apprentissage est une fonction positive et décroissante (car c'est la somme d'une fonction décroissante moins une fonction croissante) ce qui indique également un surapprentissage qui diminue quand le ration n/p augmente.

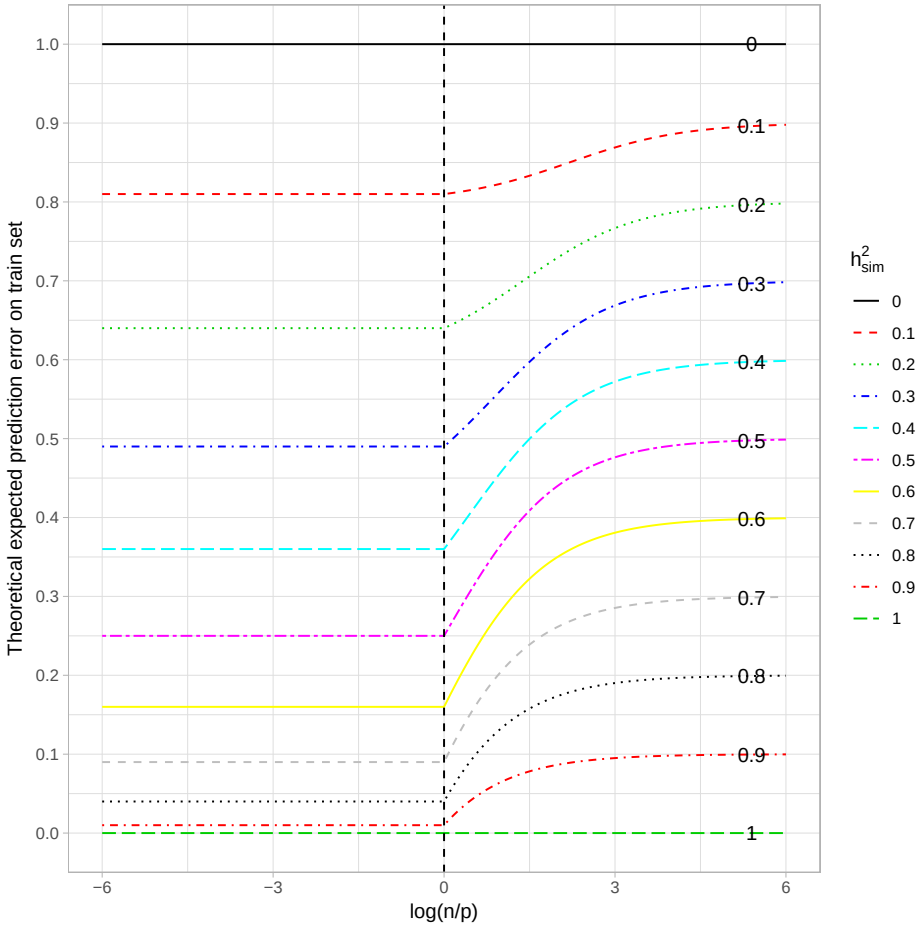


FIGURE 5.2 – Erreur quadratique théorique avec notre approximation sur l'ensemble de d'apprentissage selon le logarithme du ratio n/p en supposant que la variance totale du phénotype vaut 1. Chaque courbe correspond à une héritabilité simulée.

5.2.4 Pour le carré de la corrélation

Rappelons qu'une mesure de pouvoir prédictif couramment utilisée est le carré de la corrélation entre une prédiction d'un individu de test et son phénotype [Daetwyler et al., 2010, Goddard, 2009]. Cette quantité intuitive a l'inconvénient de mesurer la colinéarité des prédictions avec les phénotypes sans vérifier si les quantités sont à la

même échelle. Bien qu'elle ne soit pas réellement une mesure de pouvoir prédictif, cette corrélation est très utilisée dans la littérature. En utilisant notre approximation, cette quantité devient

$$\text{corr}^2(\hat{y}_{te}, y_{te}) \simeq \begin{cases} \frac{n}{p}(h^2)^2 \text{ si } n < p \\ \frac{(h^2)^2}{\frac{p}{n}(1-h^2)+h^2}, \quad \text{sinon.} \end{cases} \tag{5.7}$$

Le graphe 5.3 montre le comportement de la corrélation au carré selon le ratio n/p pour différentes valeurs de h^2. Nous voyons d'abord que cette corrélation est une fonction croissante selon n/p, qui part de 0 quand $n/p \to 0$. On remarque également que plus l'héritabilité est élevée, plus la corrélation l'est. Comme pour l'erreur quadratique sur le test, il est donc impossible de faire de bonnes prédictions dans un contexte de grande dimension même quand l'héritabilité est élevée. De l'autre côté de la courbe, le carré de la corrélation tend vers h^2 quand $n/p \to +\infty$.

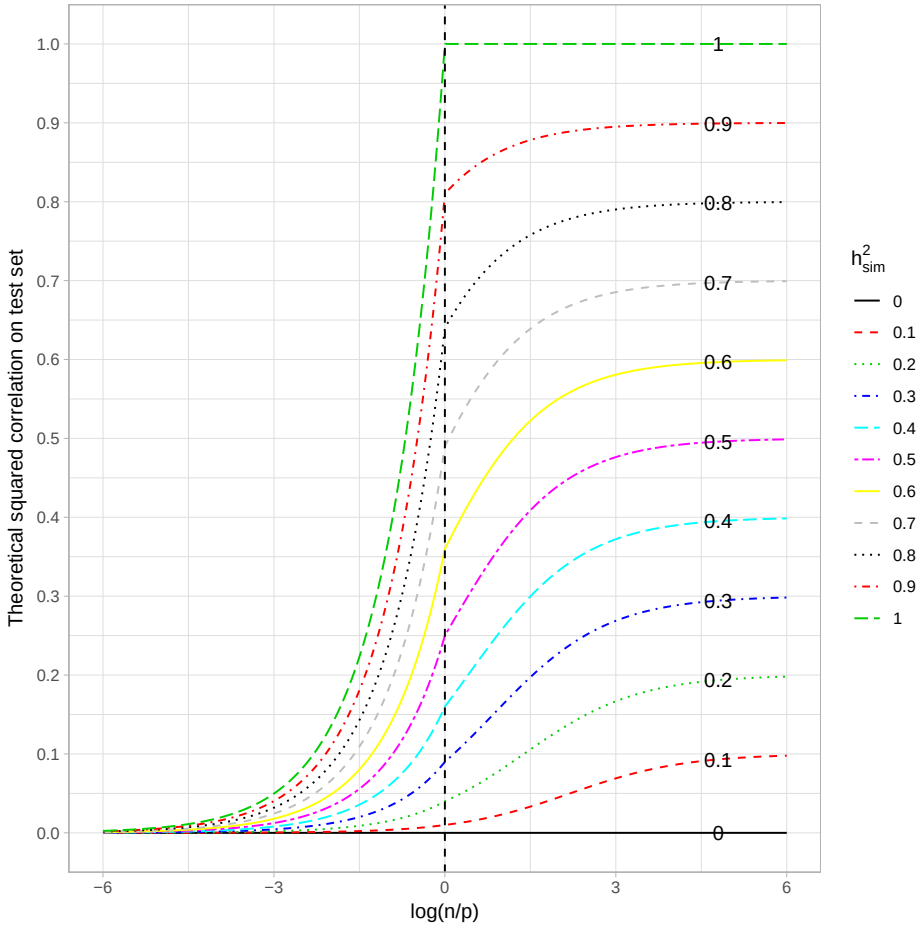


FIGURE 5.3 – Carré de la corrélation théorique avec notre approximation sur l'ensemble de test selon le logarithme du ratio n/p en supposant que la variance totale de la réponse vaut 1. Chaque courbe correspond à une héritabilité simulée.

Comme vu pour la corrélation, notre approximation peut servir à approcher diverses

quantités. Nous avons également approché le R^2 et l'héritabilité de ratio, les calculs et résultats sont disponibles en annexe.

5.3 Interprétation de l'approximation

Commençons par définir l'estimateur des effets marginaux. Pour un variant $j \in [\![1, p]\!]$, nous pouvons estimer son effet marginal selon le modèle linéaire

$$\mathbf{y} = \mathbf{z}_j u_{s,j} + \mathbf{e}.$$

L'estimateur de l'effet de ce variant par les moindres carrés est

$$\hat{u}_{s,j} = (\mathbf{z}_j^T \mathbf{z}_j)^{-1} \mathbf{z}_j^T \mathbf{y}.$$

Nous pouvons alors définir le vecteur d'estimation de l'ensemble des régressions univariées par

$$\hat{u}_s = \left(\text{diag}(\mathbf{Z}^T\mathbf{Z})\right)^{-1} \mathbf{Z}^T\mathbf{y}.$$

En supposant que $\text{diag}(\mathbf{Z}^T\mathbf{Z}) \simeq n\mathbf{I}_p$, nous avons alors

$$\hat{u}_s \simeq \frac{1}{n}\mathbf{Z}^T\mathbf{y}.$$

Notons que si nous supposons un centrage empirique, alors $\text{diag}(\mathbf{Z}^T\mathbf{Z}) = n\mathbf{I}_p$ et l'approximation ci-dessus devient une égalité.

Écrivons maintenant l'estimateur ridge avec les approximations définies plus haut. Dans le cas de la grande dimension,

$$\begin{aligned}\hat{u}_R = \mathbf{K}_\lambda \mathbf{y} &\simeq \frac{1}{p+\lambda}\mathbf{Z}^T\mathbf{y} \\ &= \frac{p}{p+\lambda}\frac{n}{p} \times \frac{1}{n}\mathbf{Z}^T\mathbf{y} \\ &\simeq \frac{n}{p}h^2\hat{u}_s.\end{aligned}$$

Notre approximation revient à estimer l'estimateur ridge par l'estimateur de l'ensemble des régressions univariées multiplié par une constante. La constante est toujours inférieure à 1 et diminue avec la "difficulté" du problème (i.e. quand le ratio n/p et l'hé-

ritabilité sont faibles). Sous cette forme on comprend la faible qualité des prédictions dans le contexte des GWAS : si n est très faible devant p alors la constante est très faible et les prédictions des effets sont très proches de 0. Nous comprenons également mieux pourquoi le terme de variance tend vers 0, puisque la constante tend vers 0.

De même en petite dimension

$$\begin{aligned}\hat{u}_R = \mathbf{K}_\lambda \mathbf{y} &\simeq \frac{1}{n+\lambda}\mathbf{Z}^T\mathbf{y} \\ &= \frac{n}{n+\lambda} \times \frac{1}{n}\mathbf{Z}^T\mathbf{y} \\ &\simeq \frac{\frac{n}{p} \times h^2}{1 + h^2 \times (\frac{n}{p} - 1)}\hat{u}_s.\end{aligned}$$

L'estimateur de la ridge peut donc encore être approché par une multiplication par une constante de l'ensemble des régressions univariées. Une étude des capacités prédictives de la ridge, des BLUPs et de l'estimateur des marginales a été réalisée dans [Zhao and Zhu, 2019]. Un futur travail sera de comparer nos approximations aux leurs et de regarder si nos résultats respectifs sont cohérents.

5.4 Simulations

5.4.1 Description des simulations

Nous allons utiliser des simulations pour vérifier notre approximation. Nous allons regarder le comportement de l'erreur attendue de prédiction sur un ensemble de test et le carré de la corrélation pour un nombre croissant de variables dans le modèle. Les simulations suivent le protocole suivant :

Nous posons $n_{tr} = 1000$, $n_{te} = 5000$, $\mathcal{E}_p$ un ensemble de nombre de variants dans l'étude avec $p_{max} = \max(\mathcal{E}_p) = 50000$ et $h^2 = 0.6$.

1. Simulation d'un vecteur de fréquences alléliques $f \sim \mathcal{U}_{p_{max}}(0.05, 0.5)$.
2. Simulation d'un vecteur d'effets génétiques $u \sim \mathcal{N}\left(0_{p_{max}}, \frac{h^2}{p_{max}}\mathbf{I}_{p_{max}}\right)$.
3. Pour $p \in \mathcal{E}_p$,
   (a) Nous posons $\lambda_{opt} = p\frac{1-h^2}{h^2}$.

(b) Génération d'un vecteur d'effets génétiques $u^p = (u_1, ..., u_p) \times \sqrt{\frac{p_{max}}{p}}$ composé des p premières composantes de u. On a $\text{var}(u) = \frac{h^2}{p} I_p$.

(c) Simulation d'une matrice de génotypes de test $\mathbf{M}_{te}$ et $\mathbf{Z}_{te}$ sa version normalisée par f.

(d) Simulation d'un vecteur de bruit $\mathbf{e}_{te} \sim \mathcal{N}(0_{n_{te}}, (1-h^2)\mathbf{I}_{n_{te}})$ et d'un vecteur de phénotype $\mathbf{y}_{te} = \mathbf{Z}_{te} u^p + \mathbf{e}_{te}$.

(e) Pour $k \in [\![1, 300]\!]$,

i. Simulation d'une matrice de génotypes de train $\mathbf{M}_{tr,k}$ et $\mathbf{Z}_{tr,k}$ sa version normalisée par f.

ii. Simulation d'un vecteur de bruit $\mathbf{e}_{tr,k} \sim \mathcal{N}(0_{n_{tr}}, (1-h^2)\mathbf{I}_{n_{tr}})$ et d'un vecteur de phénotype $\mathbf{y}_{tr,k} = \mathbf{Z}_{tr,k} u^p + \mathbf{e}_{tr,k}$.

iii. Calcul du vecteur d'estimation des effets génétiques $\hat{u}_k^p$ (avec une régression ridge avec pour paramètre de pénalisation λ_{opt}) puis calcul d'un vecteur de prédictions pour l'ensemble de test $\hat{\mathbf{y}}_{te,k}$.

(f) Calcul de $\hat{\mathbf{g}}_p = \left(\frac{1}{300} \sum_{k \in [\![1,300]\!]} \left[\hat{\mathbf{y}}_{te,k}\right]_i\right)_{i \in [\![1, n_{te}]\!]}$.

(g) Estimation de $\text{err}_p = \frac{1}{300} \sum_{i \in [\![1,n_{te}]\!]} \frac{1}{n_{te}} \left\|\mathbf{y}_{te,k} - \hat{\mathbf{g}}_p\right\|_2^2$,
de $\text{biais}_p^2 = \frac{1}{n_{te}} \sum_{i \in [\![1,n_{te}]\!]} \left(\left[z_{te}^T u - \hat{\mathbf{g}}_p\right]_i\right)^2$
et de $\text{var}_p = \frac{1}{300} \sum_{i \in [\![1,n_{te}]\!]} \frac{1}{n_{te}} \left\|\hat{\mathbf{y}}_{te,k} - \hat{\mathbf{g}}_p\right\|_2^2$.

Rappelons que les termes "biais" et "variance" sont utilisés abusivement, car nous n'utilisons pas la formulation classique du dilemme biais-variance.

Pour les simulations, nous prendrons les valeurs de p suivantes :

$$\mathcal{E}_p = \{50000, 25000, 16667, 12500, 10000, 5000, 3333, 2500, 2000, 1667,$$
$$1429, 1250, 1111, 1000, 500, 136, 79, 56, 43, 35, 29, 25, 22, 20\},$$

ce qui équivaut à prendre un vecteur de ratio égal à

$$\mathcal{E}_{n/p} = \{0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1,$$
$$2, 7.33, 12.67, 18, 23.33, 28.67, 34, 39.33, 44.67, 50\}.$$

5.4.2 Résultat pour l'erreur de prédiction sur l'ensemble de test

Comportement moyen

Les résultats sur le comportement de l'erreur, du biais et de la variance sont présentés dans le graphe 5.4. Commençons par l'erreur quadratique (figure 5.4a) : nous voyons que l'erreur quadratique moyenne suit plutôt bien notre approximation. En particulier nous remarquons que l'erreur tend vers 1 (donc l'erreur maximale) quand $\frac{n}{p} \to 0$ i.e. en très grande dimension la régression ridge est incapable de donner des prédictions acceptables. A l'inverse quand $\frac{n}{p} \to +\infty$ l'erreur tend vers $0.4 = 1 - h^2$ i.e. l'erreur irréductible et donc la régression ridge prédit de façon parfaitement satisfaisante. Remarquons également que l'approximation semble moins bonne aux alentours de $n/p \simeq 1$ ce qui n'est pas très étonnant car c'est dans cette zone que nos approximations des matrices de covariance par leurs espérances sont les moins valides.

Regardons maintenant le carré du biais (figure 5.4b). Encore une fois le comportement moyen suit bien l'approximation : le carré du biais est une fonction décroissante tendant vers $0.6 = h^2$ quand $\frac{n}{p} \to 0$ et vers 0 quand $\frac{n}{p} \to +\infty$.

La variance (figure 5.4c) est peut-être la quantité que nous approchons le moins bien. En effet même si la courbe de notre approximation suit raisonnablement bien le comportement moyen, la variance semble toujours être sous-estimée et en particulier pour la région $\frac{n}{p} < 1$. Malgré tout, l'approximation reste plutôt bonne et le comportement moyen semble tendre à droite et à gauche vers 0.

Dispersion des estimations

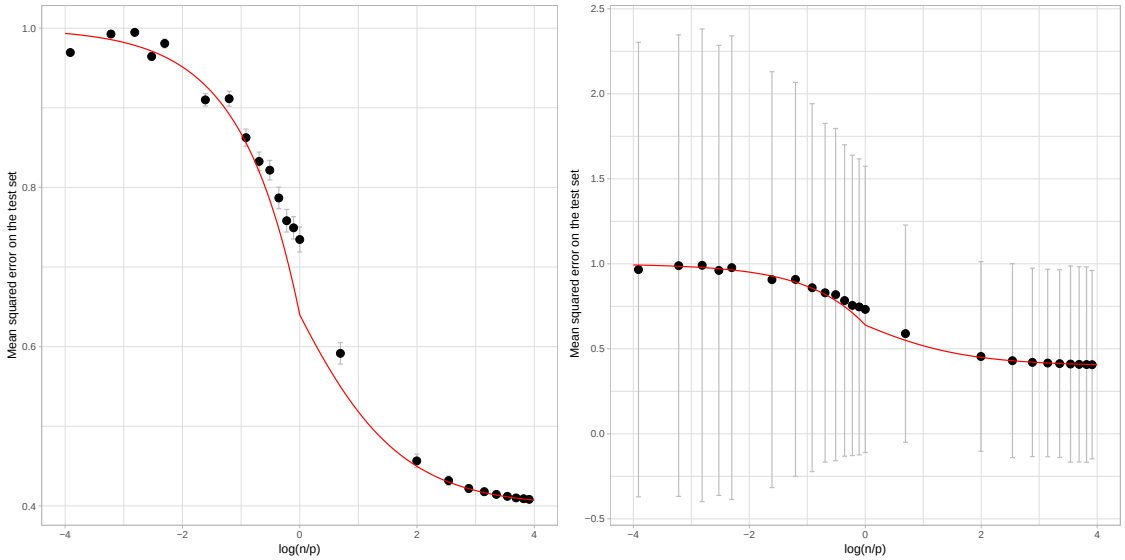
Nous avons regardé le comportement moyen de l'erreur quadratique, nous allons maintenant regarder la variabilité de ces estimations selon deux stratégies. Dans la figure 5.5a nous avons calculé l'erreur quadratique moyenne pour chacun des 300 ensembles d'apprentissage et représenté le comportement moyen plus ou moins un écart-type pour la dispersion selon ces 300 points. Dans la figure 5.5b nous avons calculé pour chaque point de l'ensemble de test une prédiction moyenne selon les 300 ensembles d'apprentissage et représenté l'écart-type au sein des individus de l'ensemble de test.

Nous voyons que les barres d'erreur sont bien plus grandes dans la figure 5.5b que dans la figure 5.5a, ce qui montre que l'erreur prédictive est principalement due à la

(a) Erreur (b) Carré du biais (c) Variance

FIGURE 5.4 – Graphe du comportement moyen de l'erreur quadratique, du carré du biais et de la variance selon le log du ratio n/p obtenus par simulations. Les points noirs correspondent aux valeurs des quantités moyennes selon les ensembles d'apprentissage et de test pour les différentes valeurs de ratio n/p. La courbe rouge correspond à notre approximation.

prédiction individuelle et peu au choix de l'ensemble d'apprentissage. Une explication pourrait être la présence dans l'ensemble de test d'individus avec un grand terme environnemental (au sens de Yang et al. i.e. avec un grand terme de bruit) : ces individus seront toujours mal prédits (même dans un cadre de petite dimension) quel que soit l'ensemble d'apprentissage, et ils font donc exploser l'erreur.



(a) Selon l'ensemble d'apprentissage (b) Selon les individus de test de test

FIGURE 5.5 – Graphe du comportement moyen de l'erreur quadratique et de sa variabilité selon le log du ratio n/p obtenus par simulations. Les points noirs correspondent au valeurs des quantités moyennes selon les ensembles d'apprentissage et de test pour les différentes valeurs de ratio n/p. La courbe rouge correspond à notre approximation. Les barres d'erreur dans les panneaux (a) et (b) correspondent à un écart-type selon deux stratégies. Le panneau (a) correspond à la variabilité de l'erreur sur les 300 ensembles d'apprentissage et le panneau (b) la variabilité des erreurs de l'ensemble de test.

5.4.3 Résultats pour le carré de la corrélation

Dans la section précédente nous avons pu vérifier la validité de notre approximation avec l'erreur prédictive. Pour pouvoir nous comparer à la littérature, nous allons maintenant regarder le comportement du carré de la corrélation sur les simulations. Notre objectif n'est pas de réaliser une comparaison exhaustive des méthodes de la littérature, mais plutôt de vérifier la validité de notre approximation.

La figure 5.6 montre le comportement du carré de la corrélation pour différents ratios n/p. Nous y voyons que la corrélation est une fonction croissante en n/p, qui tend vers 0 quand $n/p \to 0$ et également vers $h^2_{sim} = 0.6$ quand $n/p \to +\infty$. Ce n'est pas un résultat inattendu : ce sont les comportements asymptotiques de notre approximation.

Les points saumons représentant les 300 réplications de l'ensemble d'apprentissage pour chaque valeur de rapport n/p sont peu dispersés. Nous pouvons donc en conclure que comme pour l'erreur prédictive la variance inter-étude de notre quantité n'est pas très élevée.

La courbe de notre approximation (courbe verte) suit plutôt bien les points moyens. Nous remarquons encore une fois que l'approximation est moins bonne quand le ratio n/p est d'environ 1. La courbe rouge correspond à la fonction $n, p, h^2 \mapsto \left(\frac{h^2}{\sqrt{h^2+p/n}}\right)^2$ et a été proposée par Daetwyler et al. [2008] tandis que la courbe en bleu correspond à la fonction $n, p, h^2 \mapsto h^2 \times \frac{h^2}{\frac{p}{n}(1-h^2)+h^2}$ qui correspond à l'approximation de Rabier et al. [2016].

Pour la zone où $n < p$, l'approximation de Daetwyler a tendance à sous-estimer la corrélation et qu'à l'inverse notre approximation surestime la corrélation. Les deux approximations semblent toutefois être assez proches l'une de l'autre, en particulier quand $n/p \to 0$. L'approximation de Rabier semble par contre sévèrement sur-estimer la corrélation.

Dans la zone $n > p$ notre approximation et celle de Rabier sont confondues. Ici encore notre approximation surestime légèrement l'erreur et celle de Daetwyler la sous-estime. Toutefois sur cette section notre approximation semble mieux coller les simulations que celle de Daetwyler.

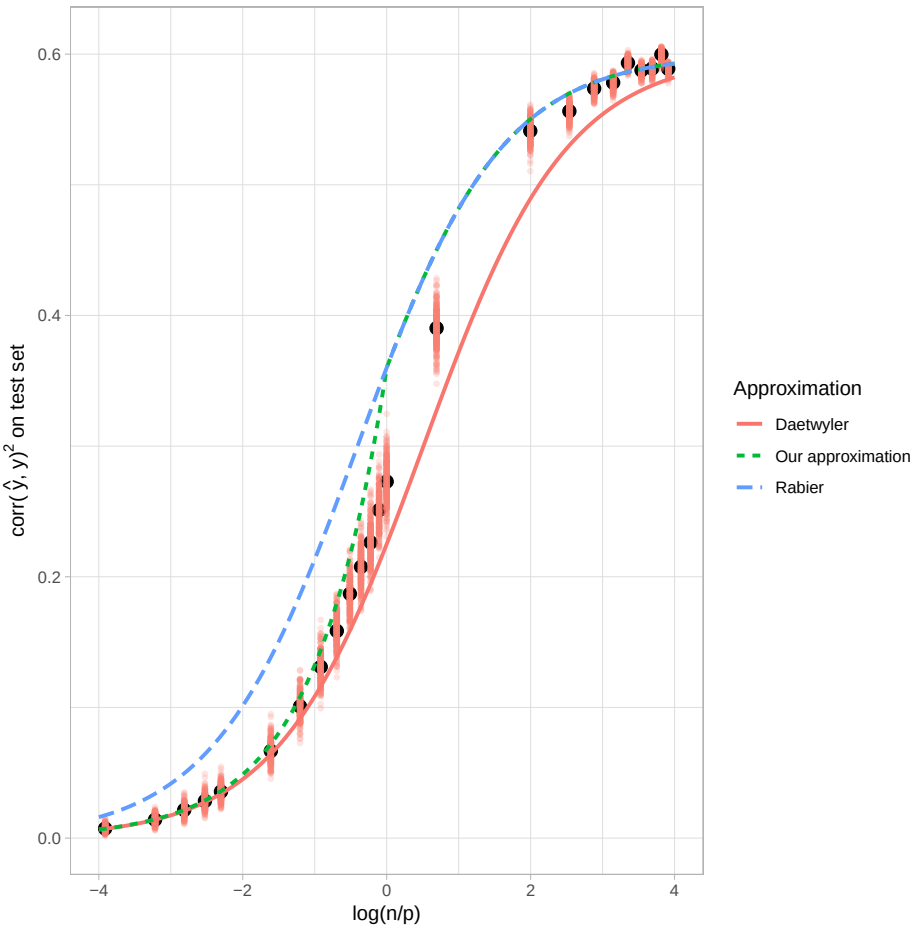


FIGURE 5.6 – Graphe du carré de la corrélation moyen sur un ensemble de test selon le ratio n/p sur des données simulées. Chaque point saumon correspond au carré de la corrélation pour une des 300 études simulées et les points noirs correspondent aux valeurs moyennes pour chaque valeur du ratio n/p simulée. L'héritabilité simulée vaut 0.6. Les courbes verte, rouge et bleu correspondent respectivement à notre approximation, celle de Daetwyler et celle de Rabier avec une héritabilité fixée à 0.6.

5.5 Application aux données UK Biobank

5.5.1 Description de l'approche

Dans les sections précédentes de ce chapitre, nous avons étudié le comportement théorique de pouvoir prédictif de la régression ridge selon le ratio n/p. Nous allons maintenant regarder si ce comportement se vérifie sur des données réelles. Sur des données réelles, nous souhaiterons souvent ajouter au modèle des variables non-pénalisées comme des covariables cliniques par exemple. Notre modèle sera de la forme

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} = \mathbf{f} + \mathbf{g} + \mathbf{e}.$$

De nombreuses questions sont ouvertes : comment calculer les différents termes de la prédiction ? comment les évaluer ? lesquels évaluer ?

Commençons par savoir comment les calculer : en effet nous voulons ici utiliser la régression ridge pour estimer les effets génétiques mais notre modèle intègre également des covariables non-génétiques. Or nos calculs des sections précédentes n'avaient pas pris en compte la présence d'effets fixes. L'estimation jointe des effets des variables non-pénalisées et pénalisées dans la régression ridge est quelque chose d'assez mal défini. Faute de mieux nous avons opté pour une estimation en deux temps des effets pénalisés et non-pénalisés. Dans le chapitre 7 nous présenterons quelques idées pour une estimation plus satisfaisante des effets.

Comme montré précédemment, en utilisant une approche basée sur une projection, il est possible d'estimer u en prenant en compte les effets fixes. Pour l'estimation des effets fixes, nous pouvons nous demander si nous devons prendre en compte le terme génétique. Devons-nous effectuer une régression linéaire sur les covariables en utilisant comme réponse les phénotypes directement (estimation disjointe) ou bien plutôt en utilisant les résidus des phénotypes obtenus par soustraction de l'estimation du terme génétique (estimation "semi-disjointe") ? Nous allons tester ces deux approches pour voir si des différences notables apparaissent.

Pour la question de savoir comment évaluer le pouvoir prédictif, nous nous intéresserons à 4 critères : le MSE, le coefficient de détermination R^2, la corrélation et l'héritabilité de ratio. Le MSE calcule la capacité de prédiction mais a pour principal défaut de n'être pas très comparable d'une étude ou d'un phénotype à l'autre. Les autres quantités permettront plus de comparaisons. Le carré de la corrélation entre

phénotype et prédiction n'est pas rigoureusement une capacité de prédiction mais a l'avantage d'être une quantité bornée et surtout d'être un standard de la littérature. Le R^2 sur un ensemble de test est une quantité qui permet d'illustrer la capacité de prédiction tout en rendant la comparaison possible (même si, comme on est sur un ensemble de test, on perd le côté borné entre 0 et 1). Enfin l'héritabilité de ratio sur un ensemble de test ne mesure pas la qualité de la prédiction mais plutôt l'amplitude de la prédiction par rapport à la réponse.

Enfin que regarder ? Nous pouvons nous intéresser à la prédiction du phénotype (donc avec prédiction des effets fixes ET des effets génétiques). Mais nous pouvons également être plus intéressés par les performances de prédiction du terme génétique uniquement et donc comparer prédiction génétique et résidus du phénotype après soustraction des effets fixes. Nous pouvons aussi regarder les prédictions par uniquement les effets fixes ou génétiques pour voir lesquels jouent un plus grand rôle.

Description des expériences

Ici nous décrivons notre protocole opératoire pour l'évaluation du pouvoir prédictif selon le ratio n/p. Contrairement aux simulations où nous faisions varier le ratio n/p en jouant sur p, nous allons ici faire varier n et p sera constant. Les différentes valeurs de n utilisées sont reportées dans la table 5.1.

Nous utiliserons les mêmes procédures de prétraitement et les mêmes covariables que dans les estimations d'héritabilité de la section 4.5. Nous utiliserons les VWB comme population et les CP fournies par UKBiobank pour prendre en compte la structure de population. Pour chaque valeur de n étudiée, nous allons sous-échantillonner plusieurs ensembles d'apprentissage et calculer les quantités d'intérêt pour chacun de ces sous-échantillonnages (voir table 5.2 pour le nombre d'ensembles d'apprentissage associé à chaque n).

Nous allons également sélectionner aléatoirement deux échantillons de 1000 individus indépendants parmi les individus de UKBiobank : un ensemble de *standardisation* destiné à l'apprentissage des effets non-pénalisés (nous pouvons nous permettre de réserver des données pour cette tâche car nous sommes en situation de richesse de données) et un ensemble de test pour évaluer les capacités prédictives. Notons que ces ensembles ne sont choisis qu'une fois et seront donc les mêmes pour toutes les tailles d'ensemble d'apprentissage et leurs répétitions respectives.

En pratique, pour chaque valeur de n et chaque répétition, notre plan d'expérience est le suivant :

1. Échantillonage d'un ensemble d'apprentissage de taille n indépendant des ensembles de standardisation et de test.

2. Application des filtres de prétraitement décrits dans la section 4.5.2 à l'ensemble d'apprentissage. Les variants exclus de l'ensemble d'apprentissage le sont également des ensembles de standardisation et de test.

3. Calcul d'un estimateur des effets génétiques avec une approche type projection pour prendre en compte les effets non-pénalisés. Le paramètre de pénalisation est obtenu par GCV.

$$\hat{u}_R = \mathbf{Z}_{tr}^T \mathbf{C}_{tr}^T \left(\mathbf{C}_{tr} \mathbf{Z}_{tr} \mathbf{Z}_{tr}^T \mathbf{C}_{tr}^T + \lambda_{GCV} \mathbf{I}_{n-r} \right)^{-1} \mathbf{C}_{tr} \mathbf{y}_{tr}$$

4. Calcul d'un estimateur des effets non-pénalisés. Nous testerons les deux approches : en utilisant comme réponse les résidus du phénotype après soustraction de l'estimation du terme génétique

$$\hat{\beta}_2 = \left(\mathbf{X}_{std}^T \mathbf{X}_{std} \right)^{-1} \mathbf{X}_{std}^T \left(\mathbf{y}_{std} - \mathbf{Z}_{std} \hat{u}_R \right). \tag{5.8}$$

   ou bien directement le phénotype

$$\hat{\beta}_1 = \left(\mathbf{X}_{std}^T \mathbf{X}_{std} \right)^{-1} \mathbf{X}_{std}^T \mathbf{y}_{std} \tag{5.9}$$

5. Calcul de l'estimateur du terme génétique, du terme non-pénalisé et de l'estimateur du phénotype sur le jeu de *test*.

$$\begin{aligned} \hat{\mathbf{g}}_{te} &= \mathbf{Z}_{te} \hat{u}_R \\ \hat{\mathbf{f}}_{te} &= \mathbf{X}_{te} \hat{\beta} \\ \hat{\mathbf{y}}_{te} &= \hat{\mathbf{f}}_{te} + \hat{\mathbf{g}}_{te} \end{aligned}$$

6. Calcul du MSE, du carré de la corrélation, du R^2 et de l'héritabilité de ratio entre les différentes quantités d'intérêt ($\hat{\mathbf{g}}_{te},\ \hat{\mathbf{f}}_{te},\ \hat{\mathbf{y}}_{te}$ et $\mathbf{y}_{te}$).

Ensemble	Cardinal
Apprentissage	$\{1000, 2000, 5000, 10000, 20000\}$
Standardisation	1000
Test	1000

TABLEAU 5.1 – Taille des ensembles d'apprentissage pour l'évaluation du pouvoir prédictif sur les données réelles.

Taille de l'ensemble d'apprentissage	1000	2000	5000	10 000	20 000
Nombre de répétitions	100	70	50	20	10

TABLEAU 5.2 – Nombre de répétitions pour l'évaluation du pouvoir prédictif sur données réelles. Valeurs à vérifier

5.5.2 Évolution des pouvoirs prédictifs selon n sur UKBiobank.

L'ensemble des résultats que nous allons maintenant commenter se résume en série de graphes. Il y a une boite de graphes pour chacun des quatre phénotypes (taille, IMC, circonférence des hanches et tour de taille). Dans chacune de ces boites se trouveront plusieurs boxplots associés à différentes quantités pour différentes mesures.

Commençons par le graphe 5.7. Dans chaque boite on trouve un boxplot d'estimation d'héritabilité (calculée en transformant le paramètre de pénalisation d'une régression ridge choisi par GCV en utilisant une matrice de contraste), un boxplot de degrés de liberté effectifs tels que définis en (1.15) divisés par la taille de l'échantillon de test (que nous appellerons d.d.l.e.n. pour *degrés de liberté effectifs normalisés*) et un boxplot de la variance empirique des coefficients de l'estimation du vecteur d'effets des variants par la régression ridge $\hat{u}_R$. Les d.d.l.e. sont une mesure de complexité classique pour la régression ridge et sont bornés entre 0 et n car nous sommes en grande dimension. En divisant ces d.d.l.e. par leur maximum, nous obtenons donc une mesure de complexité comparable pour les différentes valeurs de n.

Pour tous les phénotypes, l'estimation d'héritabilité est très dispersée quand n/p est faible puis se stabilise quand n/p augmente. En supposant que la valeur stabilisée est la plus fiable, nous obtenons ainsi une estimation à environ 0.73 pour la taille, 0.33 pour l'IMC, 0.3 pour la circonférence des hanches et 0.28 pour le tour de taille. Nous observons un comportement très similaire pour les d.d.l.e.n. : une grande instabilité pour les faibles valeurs de n, puis une stabilisation quand n augmente. Ainsi pour la taille, les d.d.l.e.n. vont tendre vers 0.66 pour la taille, 0.30 pour l'IMC, 0.27 pour la circonférence des hanches et 0.25 pour le tour de taille.

Pour la variance empirique des coefficients de l'estimation du vecteur d'effet des variants, nous constatons pour tous les phénotypes une augmentation avec n. Nous regardons si cette augmentation est linéaire en utilisant la corrélation entre le ratio n/p et cette quantité. La corrélation donne des valeurs élevées pour les différents phénotypes (plus de 90% pour la taille, environ 70% pour l'IMC et la circonférence des hanches et enfin environ 65% pour le tour de taille). Nous proposons également en annexe C.4 une approximation de la variance des coefficients du vecteur d'estimation des effets génétiques qui confirme notre intuition d'une augmentation de cette variance selon n.

Les graphes 5.8, 5.10, 5.11 et 5.9 présentent respectivement les mesures de MSE, de R^2, d'héritabilité de ratio et de corrélation phénotype/prédiction. Pour chacune de ces quantités et pour chaque phénotype, les informations sont organisées de la manière suivante :

— Dans le coin gauche en haut de chaque cadre se trouve la quantité calculée entre le phénotype et l'estimateur des effets fixes.

— En bas à gauche se trouve la quantité calculée entre les résidus du phénotype (après soustraction des effets fixes) et l'estimateur des effets génétiques.

— En haut à droite se trouve la quantité calculée entre le phénotype et la prédiction du phénotype (effets fixes + effets génétiques).

— En bas à droite se trouve la quantité calculée entre le phénotype et la prédiction du terme génétique.

Nous signalons également que sauf pour les calculs entre phénotype et prédiction du terme génétique, nous estimerons les effets fixes avec β_1 et β_2 comme décrit plus haut.

Commençons par décrire les résultats sur le MSE calculé entre le phénotype et l'estimateur des effets fixes. On remarque qu'en utilisant β_1 l'estimation des effets fixes est très stable selon n pour tous les phénotypes, ce qui n'est pas surprenant car nous sommes dans un cadre très favorable pour l'apprentissage (beaucoup de données pour un faible nombre d'effets à estimer). En utilisant β_2 nous voyons apparaître de la variance dans les estimations qui semble augmenter avec n. Nous constatons une augmentation des capacités prédictives pour la taille mais ce n'est pas le cas des autres phénotypes pour lesquels le MSE semble diminuer avec n puis augmenter pour $n = 20000$. De notre point de vue ce phénomène de remontée est un bruit statistique dû au faible nombre de points pour $n = 20000$ et n'est donc pas très inquiétant. De plus la variance entre réplications reste très faible et les MSE restent très proches de la valeur obtenue en utilisant β_1.

Regardons maintenant les MSE entre le phénotype et sa prédiction : nous remarquons que pour tous les phénotypes le MSE diminue avec n et que l'effet semble significatif. La variance des estimations semble augmenter avec n, ce qui est probablement un effet de la diminution du nombre de réplications. Nous remarquons également que l'effet de n semble plus élevé et visible pour la taille qui a une héritabilité élevée.

Par définition les MSE entre le phénotype et sa prédiction et le MSE entre les résidus du phénotype (après soustraction des effets fixes) et l'estimateur des effets génétiques sont identiques et les résultats seront donc les mêmes que ceux décrits en haut.

Enfin le MSE entre le phénotype et la prédiction des effets génétiques donne des résultats difficiles à interpréter : en effet pour la taille le MSE ne semble pas diminuer avec n, et pour les autres phénotypes une légère baisse est visible. Nous avons deux explications : d'une part les effets fixes sont sans doute beaucoup plus forts que les effets génétiques estimés (pour la taille notamment) si bien qu'ils "cachent" ces derniers. D'autre part le nombre de réplications est trop faible pour les grandes valeurs de n et amène un bruit statistique.

Intéressons nous maintenant aux résultats pour le carré de la corrélation. Pour la prédiction avec uniquement les effets fixes, nous observons le même phénomène que pour le MSE à savoir que les estimations avec β_1 sont complètement stables et que celles avec β_2 présentent une certaine variabilité. Nous ne retrouvons encore pas tout à fait une augmentation avec n, ce qu'on explique encore par un faible nombre de réplications pour les grandes valeurs de n. Encore une fois les valeurs restent très proches entre estimations par β_1 ou β_2.

Comme pour le MSE nous retrouvons une amélioration de la corrélation en fonction de n pour les prédictions de phénotype complet qu'importe la manière d'approcher β. Nous constatons une augmentation de la variance des estimations quand n augmente, encore une fois probablement due à un faible nombre de réplications pour les fortes valeurs de n. L'augmentation du pouvoir prédictif selon n est plus élevée pour la taille, probablement du fait de la plus forte héritabilité de la taille, et semble être environ la même pour les trois autres phénotypes.

Le carré de la corrélation des résidus contre les effets génétiques va nous donner des résultats plus comparables entre phénotypes, étant donné que l'importance des effets fixes varie beaucoup entre phénotypes. Évidemment on retrouve une augmentation du R^2 selon n mais également un effet héritabilité : pour la taille la corrélation passe d'environ 0.01% quand $n = 1000$ à environ 0.095 % quand n =20000, et pour les autres phénotypes l'augmentation est moins importante (on passe d'environ 0% à un moins

de 0.012 %). Il semble donc y avoir un réel effet de l'héritabilité dans l'augmentation du pouvoir prédictif.

Contrairement au MSE, nous pouvons distinguer un effet n pour la prédiction avec uniquement le terme génétique (l'effet est faible mais bien visible).

Pour l'héritabilité de ratio et le R^2 nous retrouvons beaucoup de tendances similaires : les deux quantités sont stables pour les effets fixes (avec une légère perturbation pour β_2), et nous observons une augmentation avec n pour les prédictions complètes et les résidus. Notons que les deux quantités semblent croître avec la même tendance linéaire.

Au vu du graphe 5.7, les quantités qui mesurent la complexité d'un modèle (héritabilité et d.d.l.e.n.) semblent se stabiliser quand n augmente, ce qui est relativement cohérent : en augmentant la taille de l'ensemble d'apprentissage on diminue la variance de la pénalisation optimale.

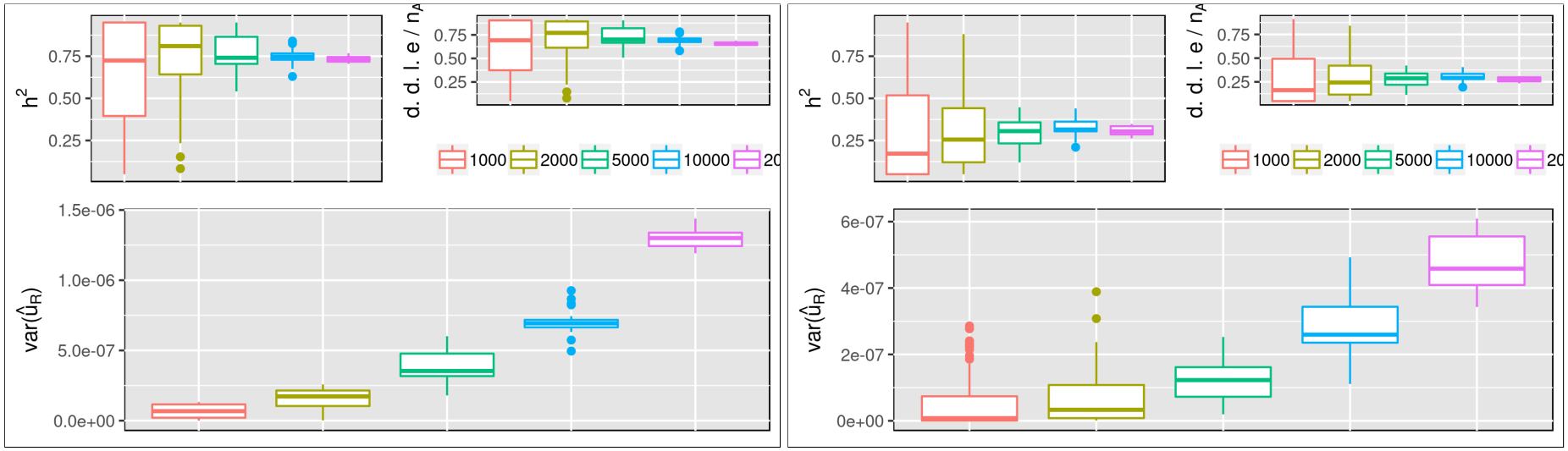
D'après les graphes 5.8, 5.9, 5.11 et 5.10 les différentes quantités semblent influencées par la taille de l'ensemble d'apprentissage (fonction décroissante pour le MSE, croissante pour les autres quantités) ce qui est satisfaisant car cela suit notre approximation. Nous en discuterons plus en détail dans la section suivante.

Un des objectifs de ces expériences était de regarder l'importance du β utilisé pour l'estimation des effets fixes. En regardant les résultats sur toutes les quantités, les phénotypes et les approches, on ne semble pas vraiment distinguer un quelconque effet. Au vu de ces résultats nous pouvons conclure que $\hat{\beta}_1$ et $\hat{\beta}_2$ donnent des résultats assez équivalents.

Notons que l'une des sources de variabilités possibles de ces expériences est le choix du paramètre de pénalisation pour la ridge. Pour retirer cette possible source de bruit nous avons refait les expériences ci-dessus en fixant l'héritabilité pour chaque phénotype pour toutes les études puis en la transformant en paramètre de pénalisation (voir la table 5.3 pour les valeurs d'héritabilité utilisées). Nous avons retracé les mêmes graphes que au dessus dans les figures C.2, C.3, C.4, C.6 et C.5. Ce changement semble avoir peu d'importance et nous retrouvons les mêmes comportements que dans les graphes que nous venons de décrire.

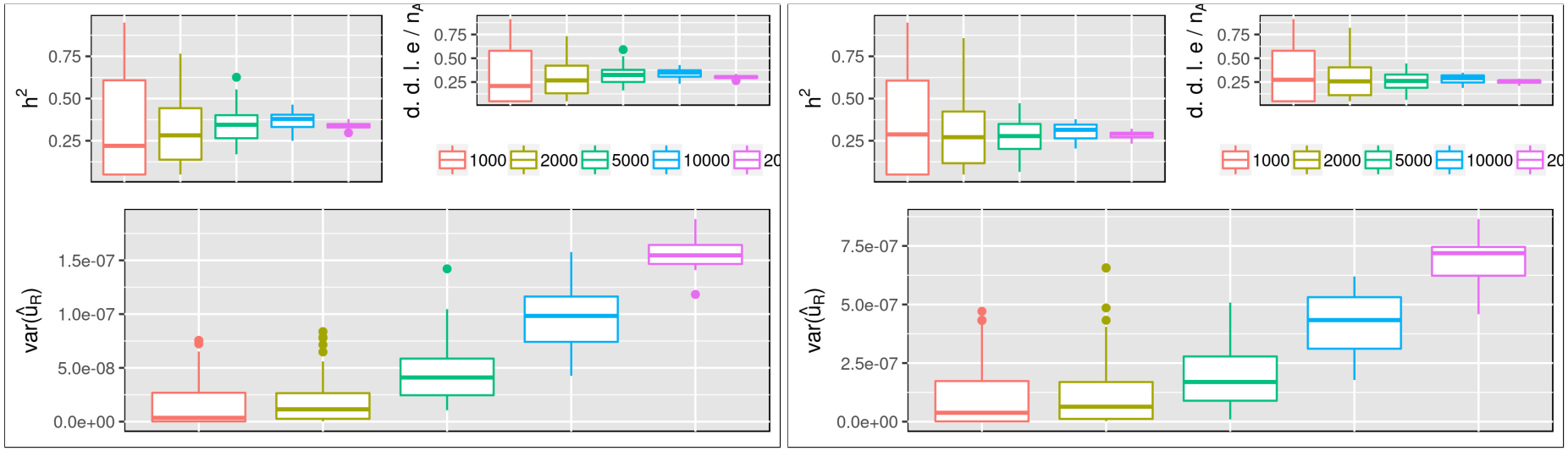
Phénotype	Taille	IMC	Circonférence des hanches	Tour de taille
Héritabilité	0.71	0.353	0.311	0.291

Tableau 5.3 – Table des valeurs d'héritabilité utilisées pour l'étude sur le comportement des quantités selon le ratio n/p et avec paramètre de pénalisation fixé.



(a) Taille

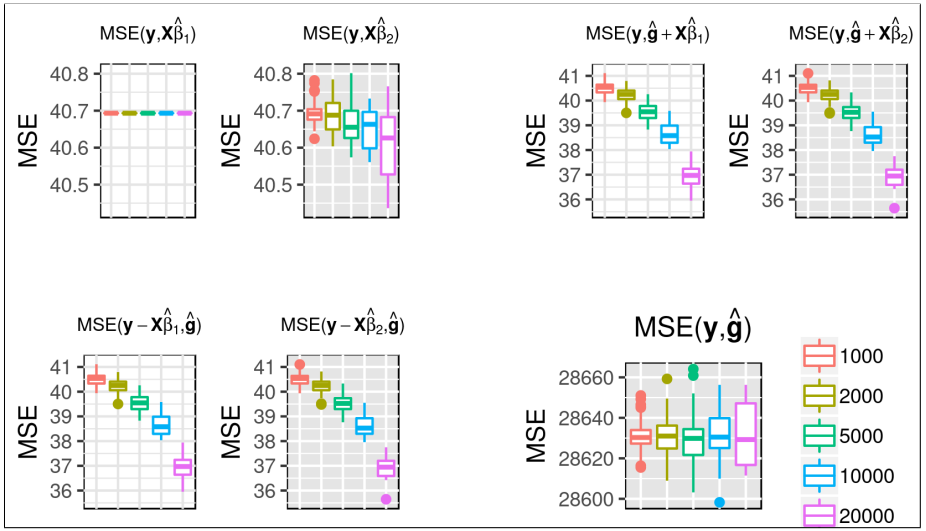
(c) Circonférence des hanches



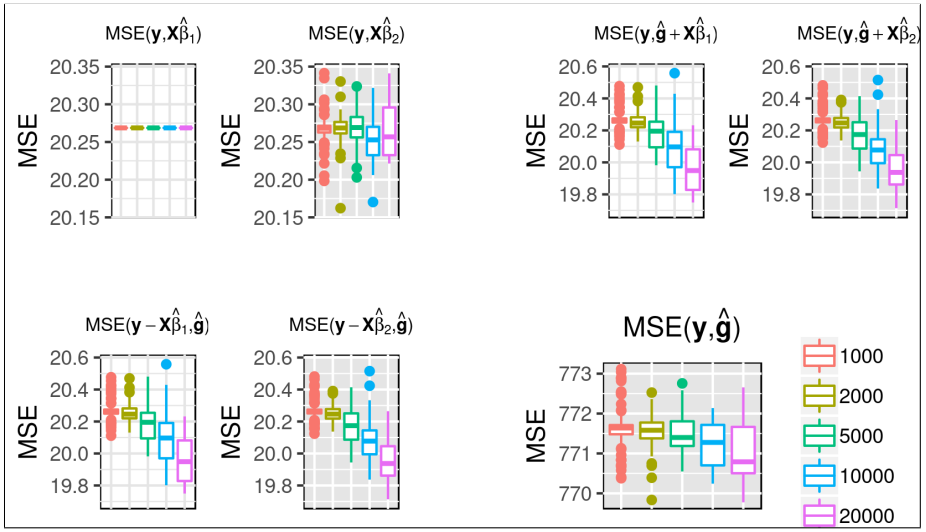
(b) IMC

(d) Tour de taille

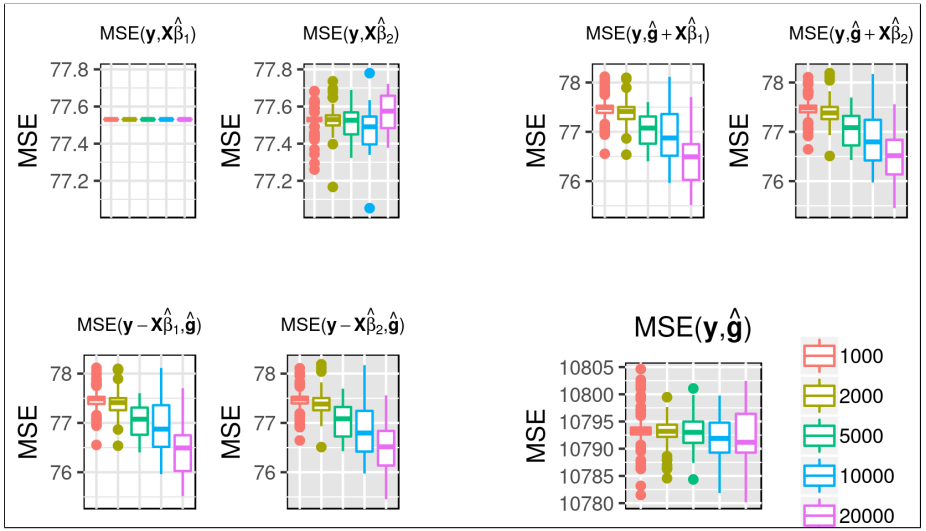
FIGURE 5.7 – Graphes d'estimation d'héritabilité, de d.d.l.e et de variance empirique de $\hat{u}_R$ sur des sous-échantillons de UKBB. Chaque boîte correspond à un phénotype. Dans chaque boîte le graphe en haut à gauche montre les estimations d'héritabilité, le graphe en haut à droite montre les degrés de liberté effectifs et le graphe du bas la variance empirique des coefficients de $\hat{u}_R$. Les quantités sont calculées pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est estimée à partir du paramètre de pénalisation de la ridge obtenu par GCV. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.



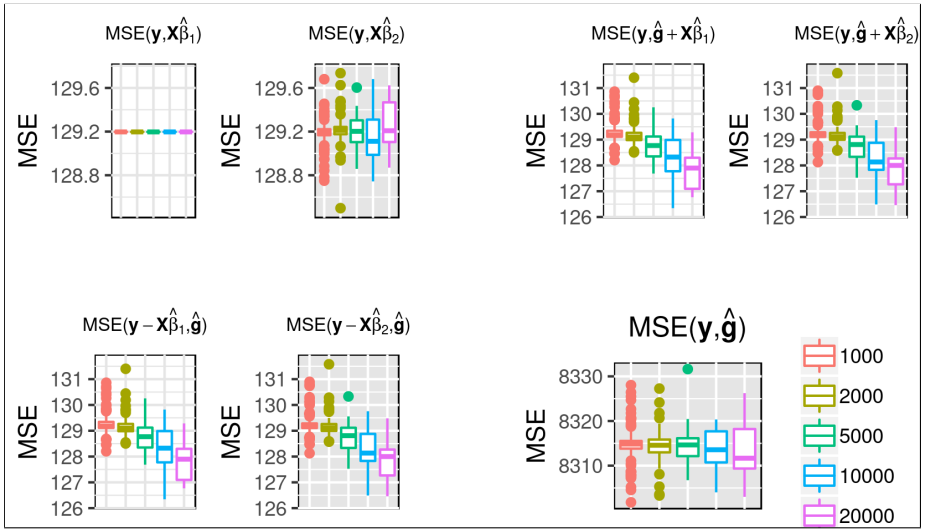
(a) Taille



(b) IMC



(c) Circonférence des hanches



(d) Tour de taille

FIGURE 5.8 – Graphes des MSE estimés sur des sous-échantillons de UKBB avec héritabilité estimée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est estimée à partir du paramètre de pénalisation de la ridge obtenu par GCV. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.

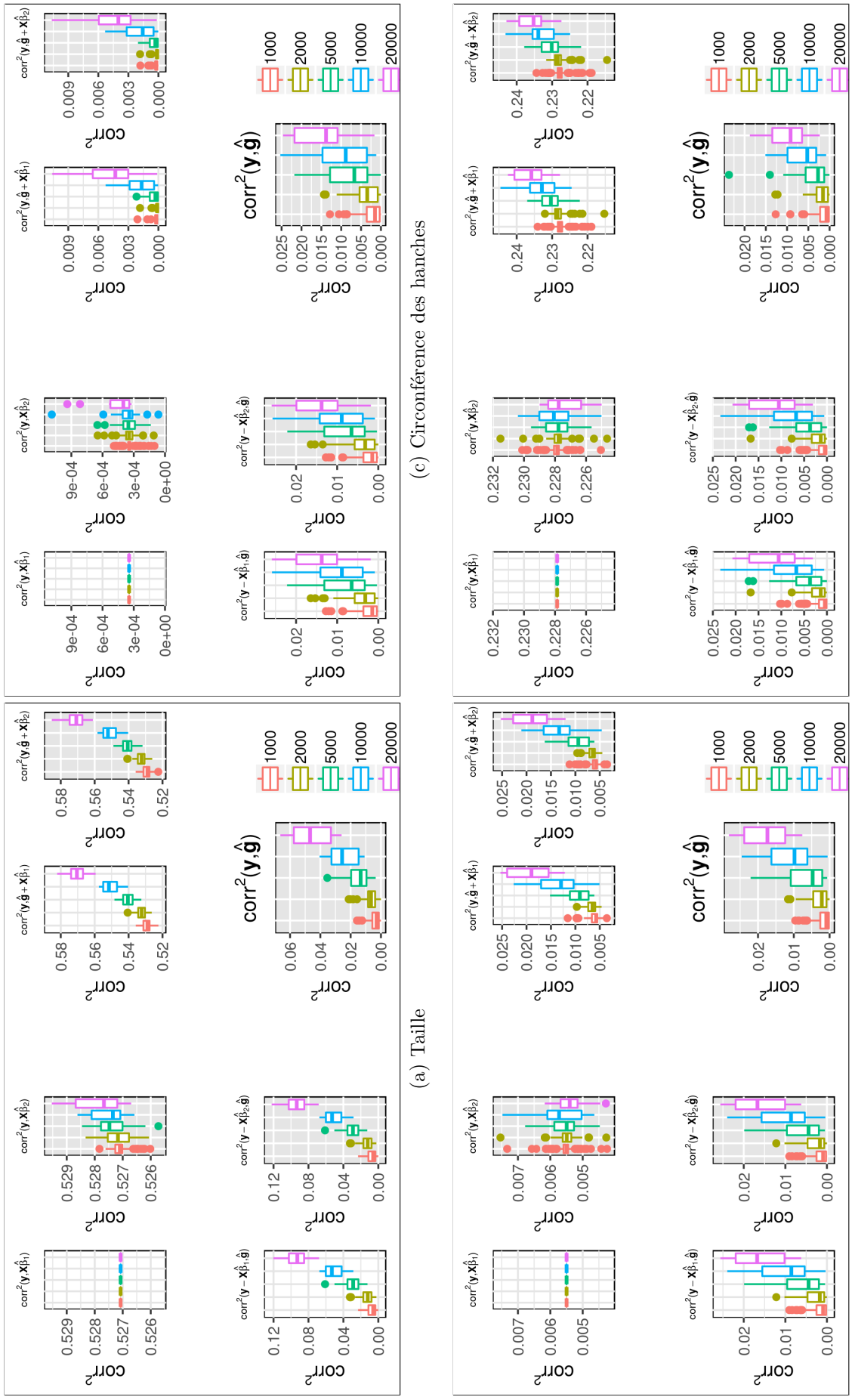
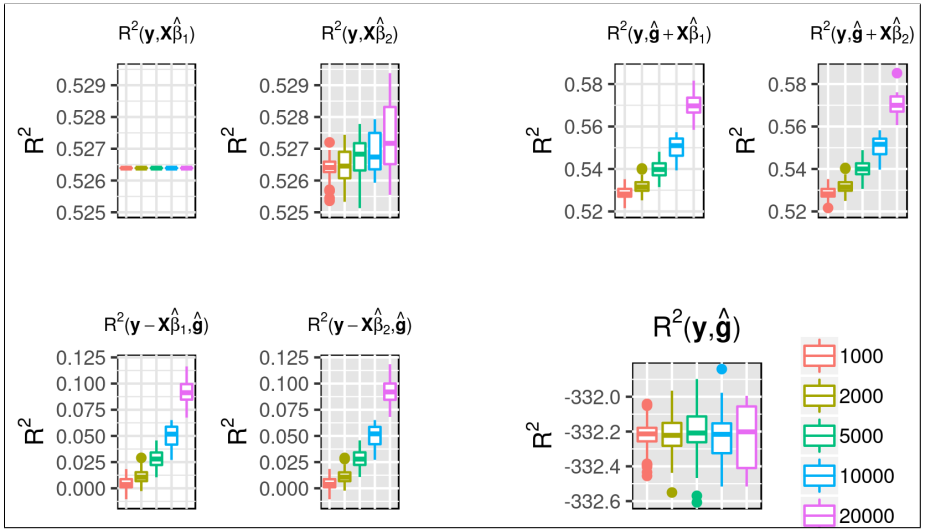
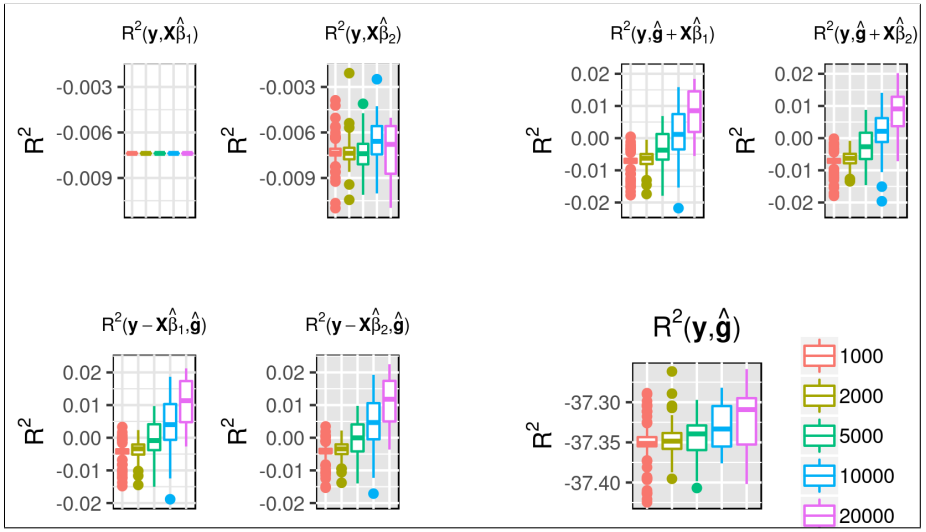


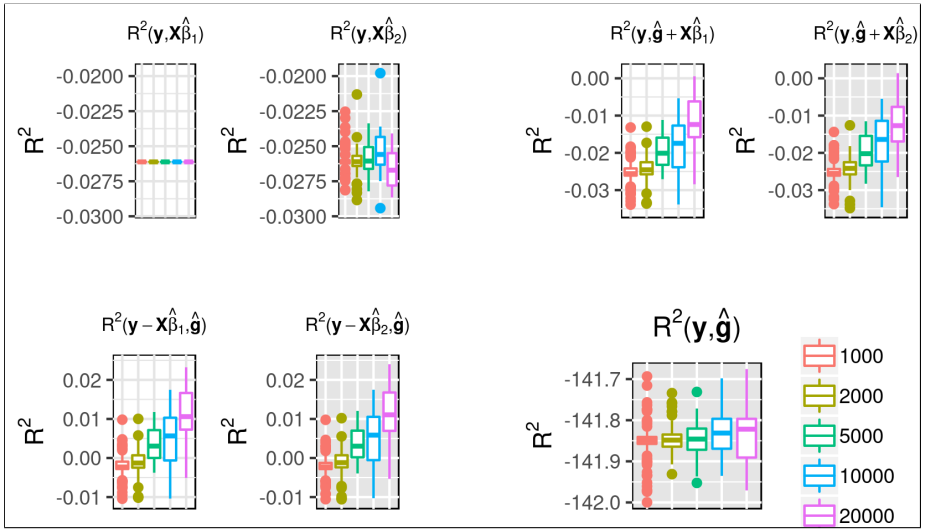
FIGURE 5.9 – Graphes des carrés de la corrélations estimés sur des sous-échantillons de UKBB avec héritabilité estimée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est estimée à partir du paramètre de pénalisation de la ridge obtenu par GCV. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.



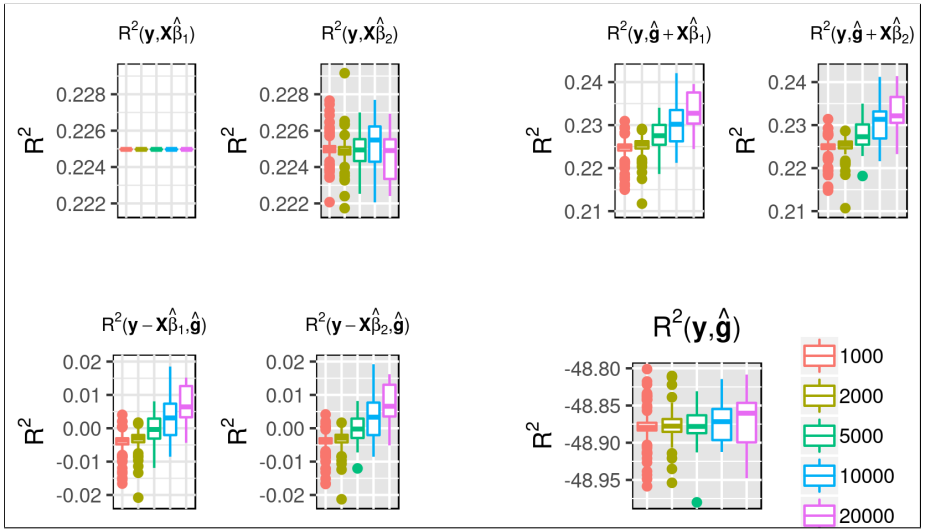
(a) Taille



(b) IMC



(c) Circonférence des hanches



(d) Tour de taille

FIGURE 5.10 – Graphes des R^2 estimés sur des sous-échantillons de UKBB avec héritabilité estimée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est estimée à partir du paramètre de pénalisation de la ridge obtenu par GCV. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.

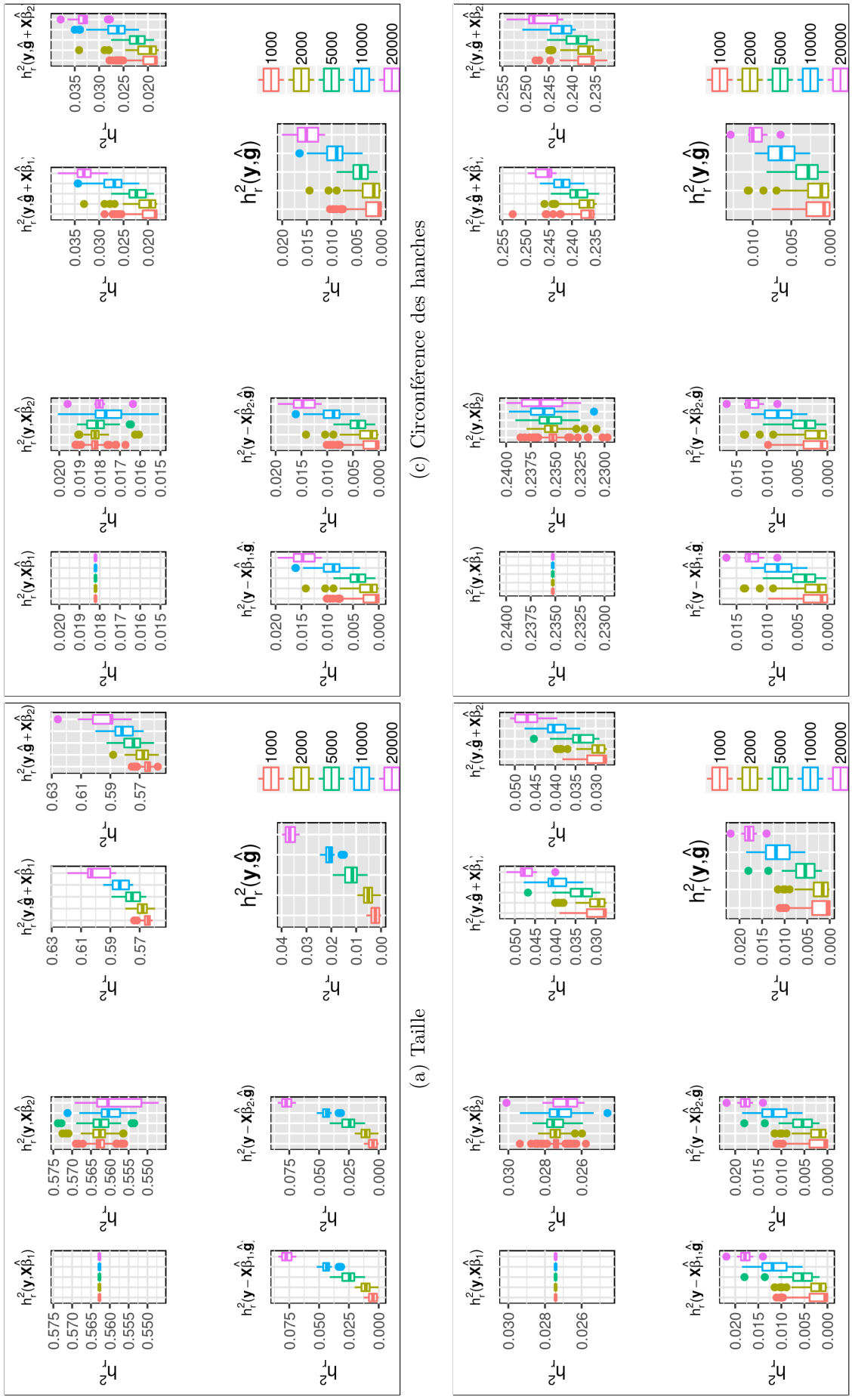


FIGURE 5.11 – Graphes des h_r^2 estimés sur des sous-échantillons de UKBB avec héritabilité estimée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est estimée à partir du paramètre de pénalisation de la ridge obtenu par GCV. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.

5.5.3 Ajustement de notre approximation aux données

Nous souhaitons voir si notre approximation est valide sur ces données réelles en nous focalisant sur l'erreur quadratique (même si nous aurions pu utiliser n'importe quelle quantité). Les principales différences entre données réelles et simulations sont la présence d'effets fixes et la perte de l'égalité $\sigma^2 = 1 - h^2$ puisque le phénotype n'est pas nécessairement de variance 1. Pour résoudre ces problèmes, nous travaillerons sur le MSE entre la prédiction génétique et les résidus des phénotypes après soustraction des effets fixes, puis nous diviserons toutes les quantités par la variance empirique calculée sur les résidus de l'ensemble de test. Pour réduire légèrement la variabilité nous allons ici supposer les héritabilités des 4 phénotypes comme des quantités fixées (en utilisant les valeurs décrites dans la table 5.3) et donc allons comparer notre approximation aux quantités décrites dans la figure C.3.

Nous avons commencé par simplement comparer dans la figure 5.12 le comportement moyen et l'approximation de l'erreur quadratique telle que nous l'avons décrite jusqu'à présent (courbe rouge). Nous voyons immédiatement que l'approximation ne suit pas le comportement moyen (particulièrement pour la taille) et semble toujours surestimer l'erreur. Ce n'est pas un résultat très étonnant car les données réelles ne suivent pas les hypothèses de notre approximation et en particulier l'hypothèse d'indépendance des variants. Le nombre de paramètres effectif sur les données réelles est donc inférieur au nombre de variants dans la régression et cela pourrait expliquer pourquoi la courbe de l'approximation est au dessus de l'erreur. Pour vérifier cette hypothèse nous allons calculer le ratio n/p "effectif" pour toutes les tailles d'ensemble d'apprentissage en utilisant la formule de notre approximation.

Pour chacun des phénotypes nous avons commencé par calculer l'erreur quadratique moyenne pour les différentes tailles des ensembles d'apprentissage puis calculé le ratio effectif moyen selon la formule

$$\mathrm{Err} = 1 - \frac{n}{p}(h^2)^2 \Rightarrow \frac{n}{p} = (1 - \mathrm{Err})\,/(h^2)^2.$$

Nous avons ensuite régressé ce ratio effectif moyen sur les tailles d'ensemble d'apprentissage associées dans la figure 5.13. Dans la figure 5.13a nous avons tracé ces points et le meilleur ajustement linéaire associé pour une régression sans intercept. Notons que toutes les valeurs des pentes sont plutôt proches les unes des autres. Nous allons ensuite utiliser la pente de cette régression comme correction pour obtenir un nombre

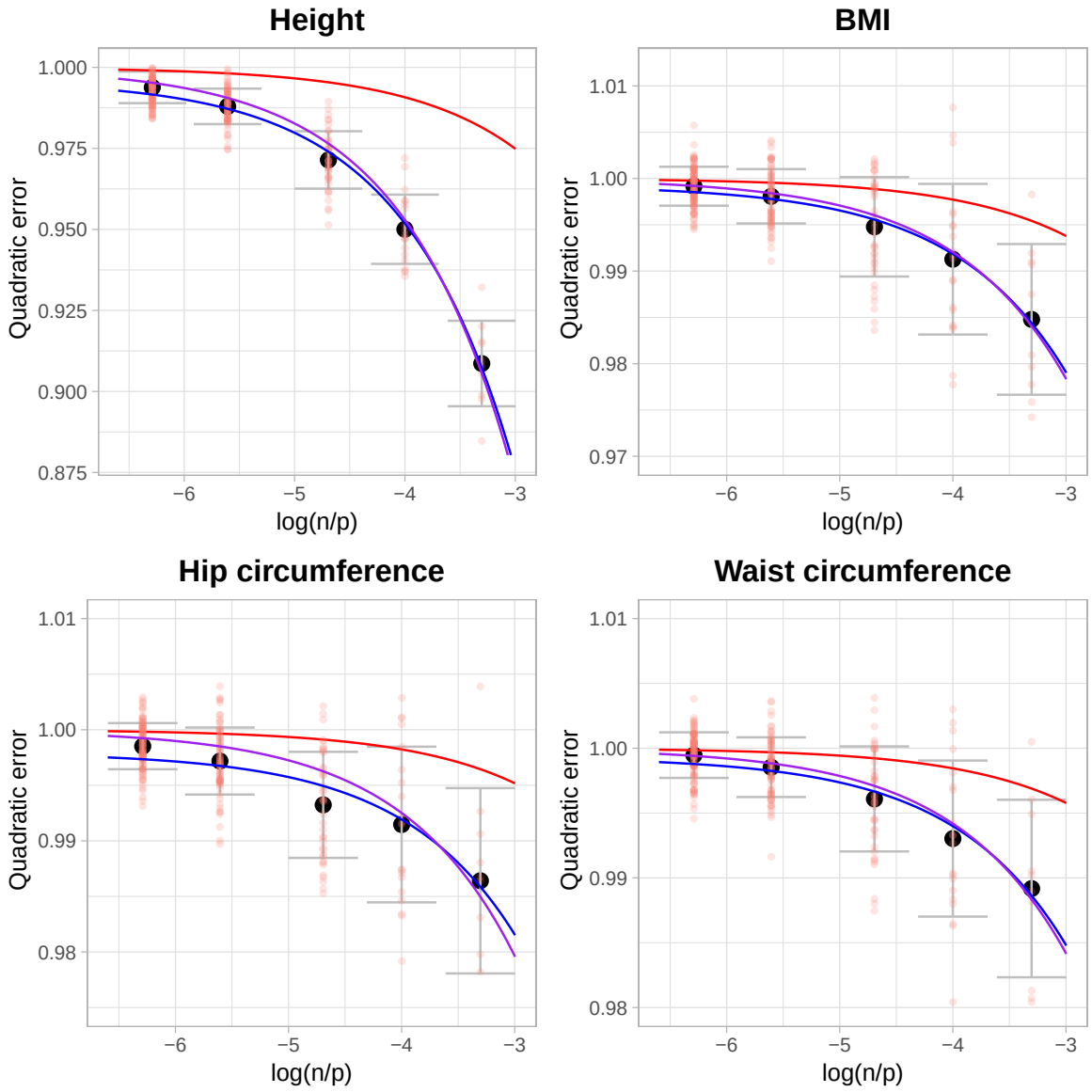


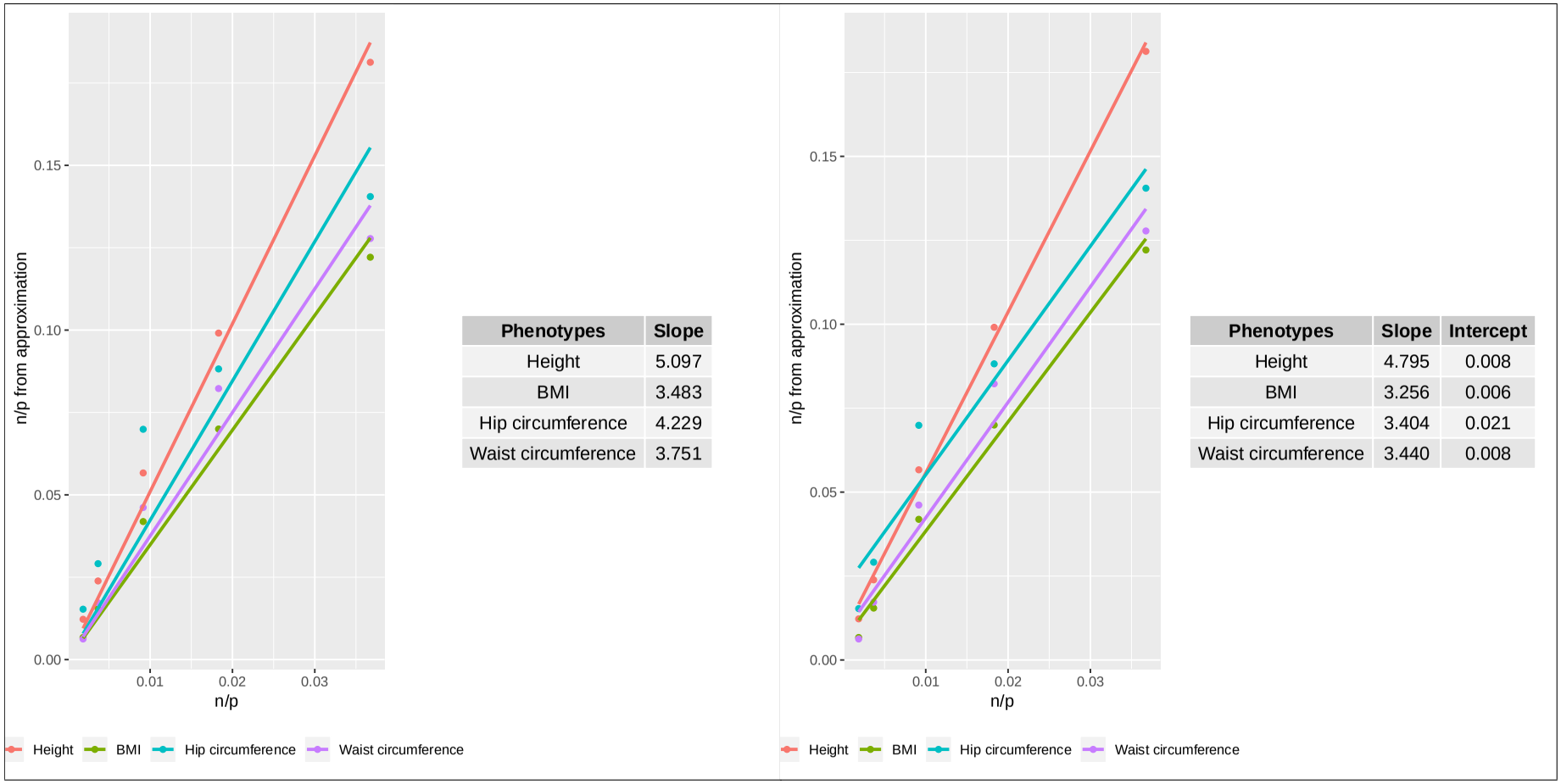
FIGURE 5.12 – Graphe de l'ajustement de notre approximation sur les données de UKBiobank. Les axes représentent le log du ratio n/p et l'erreur quadratique. Les points noirs représentent l'erreur quadratique moyenne et les points saumon l'erreur quadratique pour les différents ensembles d'apprentissage. Les barres d'erreur représentent un écart-type de la variabilité selon les ensembles d'apprentissage. Les courbes rouge, bleue et violette sont associées respectivement à notre approximation sans correction, l'approximation avec un ajustement par régression linéaire sans intercept et enfin par l'approximation avec un ajustement par régression linéaire avec intercept.

de variants effectif. La courbe bleue de la figure 5.12 correspond à notre approximation en utilisant comme valeur de p le nombre de paramètres effectif que l'on définit comme $p_{eff} = \frac{\text{Nombre de variants dans } \mathcal{A}}{\text{pente de la régression}}$. Après cette correction, nous voyons que la courbe de l'approximation colle beaucoup mieux au comportement moyen pour tous les phénotypes.

Nous avons utilisé dans la figure 5.13b la même idée d'une régression linéaire pour obtenir un nombre de variants effectif mais cette fois-ci nous ajoutons un intercept dans la régression. L'ajustement en utilisant cette régression correspond à la courbe violette dans la figure 5.12 et nous voyons que cette dernière suit également plutôt bien le comportement moyen. En particulier il ne semble pas y avoir de différences de résultats massives entre les deux approches pour estimer le nombre de paramètres effectif.

Un résultat intéressant de la régression avec un intercept est que pour les trois phénotypes dont les héritabilités estimées sont quasiment identiques (l'IMC, les hanches et le tour de taille) les pentes obtenues sont également quasiment identiques. Le gros point noir de cette approche est que nous n'avons aucune intuition pour interpréter l'intercept. Je trouve toutefois qu'il serait intéressant d'essayer cette approche sur une plus grande gamme de phénotypes avec des héritabilités estimées variées pour voir si on retrouve ce phénomène.

Un reproche que l'on peut faire à notre approche est que l'on ne teste qu'un bout de l'approximation. Nous discuterons de ce point plus en détail dans les perspectives.



Phenotypes	Slope
Height	5.097
BMI	3.483
Hip circumference	4.229
Waist circumference	3.751

Phenotypes	Slope	Intercept
Height	4.795	0.008
BMI	3.256	0.006
Hip circumference	3.404	0.021
Waist circumference	3.440	0.008

(a) Ajustement sans intercept

(b) Ajustement avec intercept

FIGURE 5.13 – Graphe des régressions linéaires pour l'ajustement de notre approximation de l'erreur quadratique aux données réelles de UKBiobank. Le panneau (a) représente un ajustement avec une régression linéaire sans intercept et le panneau (b) une régression linéaire avec intercept. A gauche de chaque panneau nous avons tracé un graphe du meilleur ajustement linéaire entre le ratio n/p de l'ensemble d'apprentissage et le ratio n/p_{eff} obtenu avec notre approximation, tandis qu'à droite se trouve un tableau avec les coefficients des régressions pour chaque phénotype.

5.6 En résumé ...

Nous avons proposé une méthode pour approximer différents pouvoirs prédictifs. Nous proposons de remplacer les matrices de ressemblances par leur espérances sous hypothèses d'indépendance (i.e. par des matrices diagonales). L'originalité de notre approximation est qu'elle est dépendante de la valeur de n/p : quand $n < p$ nous remplacerons la matrice de ressemblance des individus $\mathbf{Z}\mathbf{Z}^T$ par $p\mathbf{I}_n$ tandis que quand $n > p$ nous remplacerons la matrice de ressemblance des variants $\mathbf{Z}^T\mathbf{Z}$ par $n\mathbf{I}_p$. Nous pouvons alors utiliser ces approximations pour le MSE (sur ensemble de test et d'apprentissage) et pour le carré de la corrélation (mais également le R^2 statistique et l'héritabilité de ratio).

Un travail avec des simulations nous a permis de montrer la validité de notre approximation (pour toutes les quantités), en particulier dans les zones $n << p$ et $n >> p$: l'approximation suivait plutôt bien les quantités calculées sur les simulations.

Nous nous sommes également intéressés à une application sur UK-Biobank. Un ajustement de l'approximation s'est révélé nécessaire pour intégrer le déséquilibre de liaison de ces données. Après ajustement, l'approximation est plutôt satisfaisante.

Chapitre 6

L'estimation d'héritabilité pour les phénotypes qualitatifs

Dans cette section nous présenterons quelques résultats sur l'estimation d'héritabilité appliquée aux maladies et plus généralement aux phénotypes binaires. Cette section s'est beaucoup appuyée sur le travail de stage de fin d'études de Mathilde Carlier. Au bout d'environ un an de thèse, nous avons décidé de basculer sur le cas quantitatif car l'étude de l'estimation d'héritabilité pour le cas binaire s'est révélée assez délicate. Nous avions pour objectif de revenir sur le cas binaire, mais nous n'avons malheureusement pas eu le temps. De plus nous n'avions pas encore travaillé sur la régression ridge à ce moment là donc toutes les méthodes d'estimation utiliseront le modèle mixte.

Nous présenterons dans cette section le modèle de la liability et deux méthodes existantes pour l'estimation d'héritabilité pour les phénotypes qualitatifs. La première utilise les modèles à effets aléatoires pour estimer les composantes de variance, avec une correction pour intégrer les spécificités du cas binaire. Les estimations d'héritabilité données par cette approche présentent des biais que nous exposerons. La deuxième propose une estimation directe de l'héritabilité à l'aide d'une régression entre produit de phénotypes et ressemblance génétique et se veut sans biais. Nous discuterons des biais de la première approche, montrerons la sensibilité de ces deux approches aux filtres de prétraitement et proposerons une explication pour les valeurs aberrantes parfois renvoyées par ces méthodes.

6.1 Contexte

6.1.1 Modèle de liability et calculs de Falconer

Les calculs de l'héritabilité de Fisher ne s'intéressaient qu'aux phénotypes quantitatifs et ne peuvent donc pas être étendus directement aux phénotypes qualitatifs que sont les maladies. Falconer [2007] propose une méthode pour estimer l'héritabilité pour des maladies à partir de données familiales. Sa méthodologie s'appuie sur le *Liability Threshold Model* : il suppose que pour chaque phénotype binaire il existe un phénotype sous-jacent continu et gaussien appelé *liabilité* (ou *liability* en anglais). Si la liability d'un individu est supérieure à un certain seuil t, alors cet individu sera considéré comme un cas et inversement. Le principe du modèle de liability est présenté graphiquement dans la figure 6.1. Les calculs de Falconer permettent d'arriver à une estimation de l'héritabilité pour la liability (on parle d'*héritabilité à l'échelle de la liability* h_l^2) et on utilisera cette valeur comme héritabilité du trait binaire. Nous ne détaillerons pas ici les calculs de Falconer et donnerons simplement le résultats de son calcul d'h_l^2.

$$h_l^2 = \frac{2\left(t - t_1\sqrt{1-(t^2-t_1^2)(1-(t/a^2))}\right)}{a + t_1^2(a-t)} \tag{6.1}$$

avec

- t le seuil de liability au delà duquel un individu est malade : $l > t \Rightarrow y = 1$,
- $t_1 = t - \mu_{l,R}$ avec $\mu_{l,R}$ la moyenne de la liability des parents des cas,
- a la différence entre la liability moyenne de la population générale et la liability des cas.

6.1.2 Calculs d'héritabilité sur des individus non-apparentés

Une première méthode basée sur des correctifs

Cette première méthode développée par Lee et al. [2011] utilise le principe des calculs de Falconer (i.e. passer par la liability) sur des données d'individus non apparentés. Dans leur approche, les auteurs proposent de transformer une héritabilité calculée directement sur les phénotypes binaires en héritabilité à l'échelle de la liability. Nous ne détaillerons pas les calculs des auteurs mais nous expliquerons rapidement leurs principes.

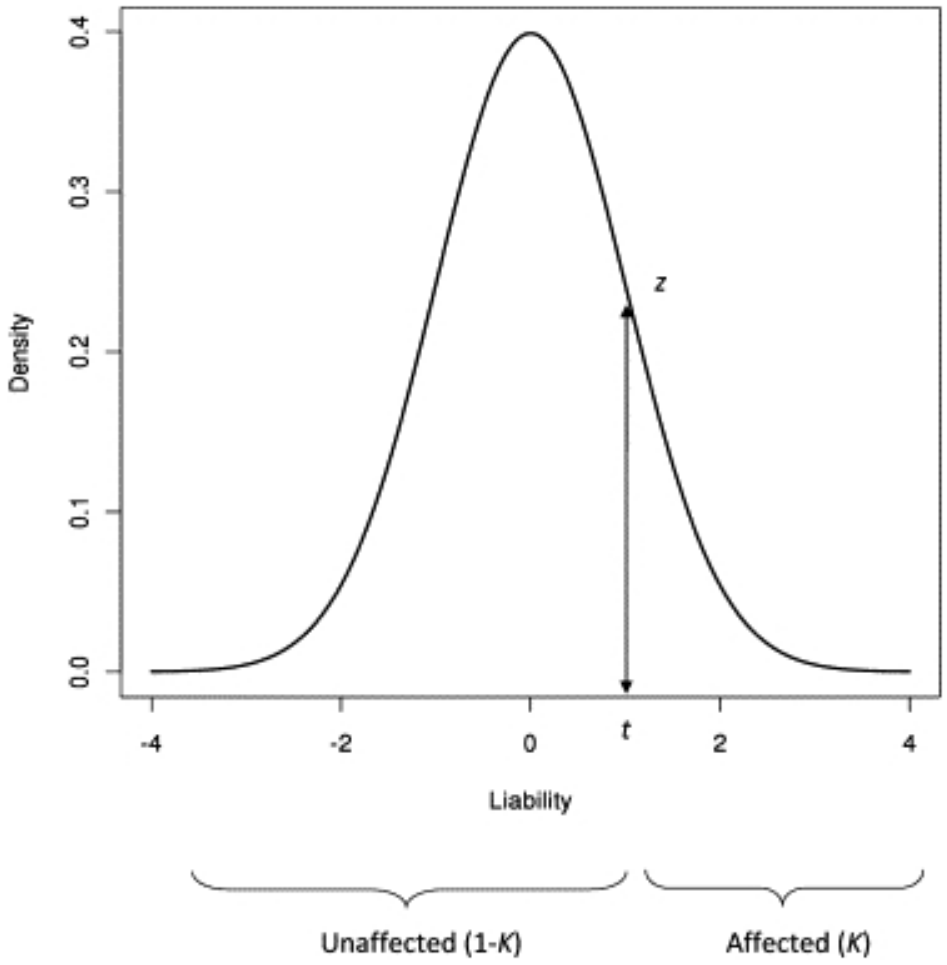


FIGURE 6.1 – Graphe du modèle de liability. La liability suit une distribution normale centrée et réduite. K est la proportion de cas dans la population. t représente le seuil de liability au delà duquel on est considéré comme un cas. z représente la hauteur de la courbe au point t. Graphe issu de Lee et al. [2011].

Commençons par poser deux modèles

$$\mathbf{y} = \mu_o \mathbb{1}_n + \mathbf{g}_o + \mathbf{e}_o, \tag{6.2}$$

$$\mathbf{l} = \mu_l \mathbb{1}_n + \mathbf{g}_l + \mathbf{e}_l. \tag{6.3}$$

Le modèle (6.2) est le modèle dit à l'échelle de l'observation : $\mathbf{y} \in \mathbb{R}_n$ est un vecteur de phénotypes binaires observés, $\mu_o \in \mathbb{R}$ la moyenne du phénotype, $\mathbf{g}_o \in \mathbb{R}_n$ un vecteur d'effets génétiques constitué de l'agrégat des effets des variants et $\mathbf{e}_o \in \mathbb{R}_n$ un vecteur d'effets environnementaux. En particulier les auteurs supposent que la matrice de covariance de $\mathbf{y}$ est de la forme $\mathbf{\Sigma}_o = \sigma^2_{\mathbf{g}_o} \mathbf{G}_* + \sigma^2_{\mathbf{e}_o}$ avec $\mathbf{G}_*$ la matrice de ressemblance génétique entre individus aux variants causaux (on suppose donc l'indépendance entre effets génétiques et environnementaux) .

Le modèle (6.3) représente lui le modèle à l'échelle de la liability. Si on suppose que la prévalence de la maladie dans notre étude est la même que dans la population générale alors la liability suit une gaussienne centrée réduite par définition $\mathbf{l} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$, $\mu_l \in \mathbb{R}$ la moyenne de la liability, $\mathbf{g}_l \in \mathbb{R}_n$ représente le terme d'effets génétiques à l'échelle de la liability et $\mathbf{e}_l$ le vecteur d'effets environnementaux à échelle de la liability. On supposera encore une fois l'indépendance entre effets génétiques et environnementaux. Nous posons une hypothèse sur les coefficients de $\mathbf{g}_l \sim \mathcal{N}(0, \sigma^2_{\mathbf{g}_l})$ et comme la variance totale de la liability vaut 1, les auteurs posent $h^2_l = \sigma^2_{\mathbf{g}_l}$.

Les auteurs définissent alors l'héritabilité à l'échelle de l'observation

$$h^2_o = \frac{\mathrm{var}(\mathbf{g}_o)}{\mathrm{var}(\mathbf{y})} = \frac{\sigma^2_{\mathbf{g}_o}}{K(1-K)}. \tag{6.4}$$

avec K la proportion de cas dans la population (voir figure 6.1).

Pour passer de l'échelle de l'observation à l'échelle de la liability, les auteurs vont utiliser un résultat de Dempster and Lerner [1950] : en écrivant la régression de $\mathbf{y}$ contre $\mathbf{l}$ les auteurs montrent que nous pouvons écrire $g_o = z g_l + c$ avec z la hauteur de la densité de la liability au point t (voir figure 6.1) et c une constante. En jonglant entre calcul sur les composantes de vecteurs et quantités définies sur les vecteurs complets, les auteurs arrivent à l'expression suivante

$$h^2_l = \frac{K(1-K)}{z^2} h^2_o. \tag{6.5}$$

En pratique ces résultats n'ont que très peu d'utilité pratique car ce calcul exigeait un hypothèse importante : la proportion de cas (appelée *prévalence*) dans notre étude K_e doit être la même que celle dans la population générale. Vu le coût du génotypage cette hypothèse ne sera jamais respectée en pratique et tout particulièrement dans le cas des maladies rares. Désormais nous utiliserons K pour parler de la prévalence de la maladie dans la population générale et K_e pour parler de la prévalence dans l'étude. Ce déséquilibre entre K et K_e pose un important problème sur nos hypothèses de normalité. En effet nous avons supposé que la distribution de la liability était normale sur l'ensemble de la population. En conséquence dans notre étude enrichie en cas, nous allons fortement perturber cette hypothèse. En effet nous aurons dans notre étude plus d'individus avec une somme effets génétiques + effets environnementaux supérieure au seuil t que dans la population générale. En conséquence la distribution des effets génétiques et environnementaux ne suit plus du tout une gaussienne. Ce phénomène est bien illustré dans le panel C de la figure 6.2 : on voit bien l'effet de l'enrichissement en cas sur les distributions de $\mathbf{g}_l$ et $\mathbf{e}_l$.

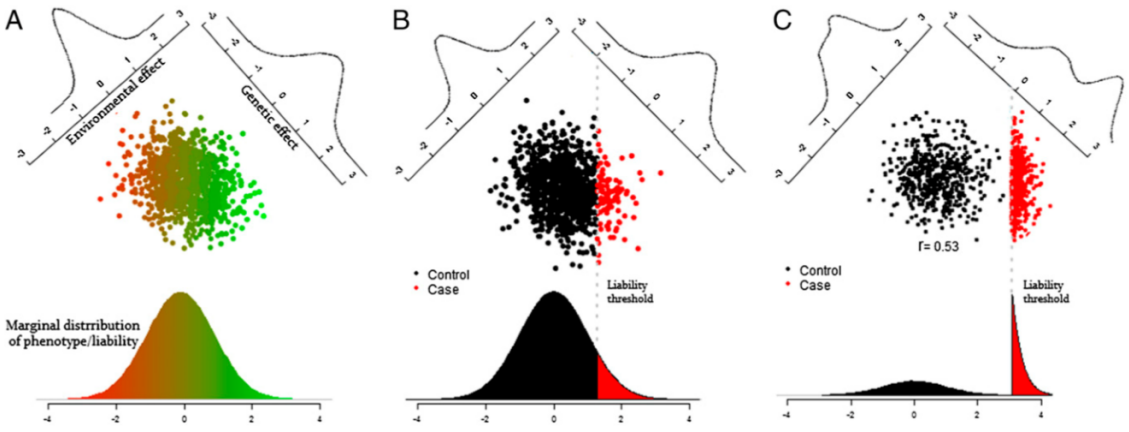


FIGURE 6.2 – Graphe sur l'influence du déséquilibre K et K_e sur la liability. Chaque panel correspond à un scénario et pour chaque panel on a en haut un graphe de dispersion de la liability selon les effets génétiques / environnementaux et en bas la distribution de la liability. Le panel A correspond au cas où le phénotype est quantitatif. Le cas B correspond au cas où le phénotype est qualitatif et $K = K_e$: toutes les liability au-dessus du seuil correspondent à des patients et les distributions des effets (génétiques et environnementaux) et donc de la liability sont gaussiennes. Dans le graphe C on est dans le cas $K \neq K_e$ et donc les hypothèses de normalité ne sont plus valides. Ce graphe est issu de Golan et al. [2014].

En supposant les hypothèses que nous avions posées pour arriver à (6.5) et en utilisant les propriétés des gaussiennes tronquées [contributors, 2020], les auteurs proposent

un calcul prenant en compte ce déséquilibre

$$h_l^2 = \frac{K^2(1-K)^2}{K_e(1-K_e)z^2} h_o^2. \tag{6.6}$$

Pour l'application pratique, les auteurs utilisent les raisonnements déjà décrits dans Yang et al. [2010], et estiment la matrice $\mathbf{G}_*$ par $\mathbf{G}$ la matrice de ressemblance génétique calculée sur l'ensemble des marqueurs génotypés. L'héritabilité à l'échelle de l'observation sera donc calculée avec des méthodes basées sur le REML puis transformée à l'échelle de la liability.

Une deuxième méthode basée sur la régression de Haseman–Elston

Golan et al. [2014] proposent une deuxième approche appellée *Phenotype Correlation–Genetic Correlation regression* (*PCGC*) pour estimer l'héritabilité à l'échelle de la liability. Leur approche se base sur une régression des produits de phénotypes contre la corrélation génétique entre individus (aussi appelée régression de Haseman–Elston Haseman and Elston [1972]). Nous ne détaillerons pas les calculs mais nous présentons ici leur raisonnement.

La base de PCGC est de supposer qu'il existe une fonction qui lie l'espérance d'un produit de phénotypes avec la matrice de ressemblance génétique $\mathbb{E}[y_i y_j] = f(h_l^2, G_{i,j}^*)$. Le cas le plus simple de cette approche est quand le phénotype est quantitatif et que l'on suppose un modèle polygénique additif. Dans ce cas $f(h^2, G_{i,j}^*) = h^2 G_{i,j}^*$ ce qui correspond à la régression de Haseman–Elston. Les auteurs proposent de déterminer une fonction f adaptée au cas binaire avec enrichissement en cas.

Posons

$$W_{ij} = \frac{(y_i - K_e)(y_j - K_e)}{K_e(1-K_e)} \tag{6.7}$$

le produit de phénotypes normalisés, et posons également et $\mathcal{S}_{ij}$ une variable d'appartenance à l'étude : $\mathcal{S}_{ij} = 1$ si les individus i, j sont compris dans l'étude et 0 sinon.

Nous souhaitons régresser W_{ij} par $\mathbf{G}_{i,j}^*$ pour les paires d'individus et nous allons donc calculer $f(h_l^2, \mathbf{G}_{i,j}^*) = \mathbb{E}[W_{ij} | \mathcal{S}_{ij} = 1, \mathbf{G}_{i,j}^*]$. En développant l'expression de l'espérance autour des trois valeurs possibles de W_{ij} et en utilisant la formule de Bayes, nous arrivons à $\mathbb{E}[W_{ij} \mid \mathcal{S}_{ij} = 1, \mathbf{G}_{i,j}^*] = \frac{\frac{1-K_e}{K_e}\mathbb{P}(Y_i=Y_j=1|\mathbf{G}_{ij}^*) - \frac{(1-K_e)K}{(1-K)K_e}\mathbb{P}(Y_i \neq Y_j|\mathbf{G}_{ij}^*) + \frac{K_e}{1-K_e}\left[\frac{(1-K_e)K}{(1-K)K_e}\right]^2 \mathbb{P}(Y_i=Y_j=0|\mathbf{G}_{ij}^*)}{\mathbb{P}(\mathcal{S}_{ij}=1|\mathbf{G}_{ij}^*)}$.

En effectuant un développement limité autour de $\mathbf{G}^*_{i,j} = 0$ (ce qui correspond à des individus non-apparentés) et en passant les phénotypes à l'échelle de la liability (qui est une variable gaussienne), les auteurs arrivent à l'expression

$$\mathbb{E}[W_{ij}|\mathcal{S}_{ij} = 1, \mathbf{G}^*_{i,j}] = \frac{K_e(1-K_e)z^2}{K^2(1-K)^2} h_l^2 \mathbf{G}^*_{i,j}. \tag{6.8}$$

Le raisonnement est détaillé en annexe D.1.

En utilisant ce résultat, les auteurs proposent une procédure en deux étapes pour estimer h_l^2 :

— Calculer $\hat{a}$ le coefficient de la régression linéaire des produits de phénotypes W_{ij} sur les coefficients de la matrice de ressemblance génétique des marqueurs $\mathbf{G}$ (qui estime la matrice de ressemblance génétique des variants causaux).

— Estimer $\hat{h}_l^2 = \hat{a}/c$ avec $c = \frac{K_e(1-K_e)z^2}{K^2(1-K)^2}$.

Nous remarquons en particulier que la constante c est la même que la constante corrective (pour intégrer le passage à l'échelle de la liability et pour prendre en contre la différence de prévalence entre la population générale et notre étude) dans l'équation (6.6). Les auteurs proposent une extension de l'approche pour l'intégration de covariables non-pénalisées et le calcul de plusieurs héritabilités pour différentes sections du génome.

6.2 Biais observés pour GCTA

Dans leur article présentant PCGC Golan et al. [2014] montrent avec des simulations des biais de GCTA dans l'estimation d'héritabilité pour des phénotypes binaires. Nous avons commencé par refaire ces simulations pour essayer de comprendre ces biais.

6.2.1 Description des simulations

Nous commençons par décrire le processus de simulation de Golan et al. [2014]. Leur modèle de simulation permet d'avoir des génotypes simulés mais surtout la prise en compte du déséquilibre entre K et K_e. Les auteurs fixent les probabilités d'être sélectionné dans l'étude à $P_{cas} = 1$ pour les cas et $P_{contrôle} = \frac{K(1-K_e)}{K_e(1-K)}$ pour les contrôles (ce qui correspond aux hypothèses de "full ascertainment"). Notons que dans ces simulations, tous les variants sont causaux.

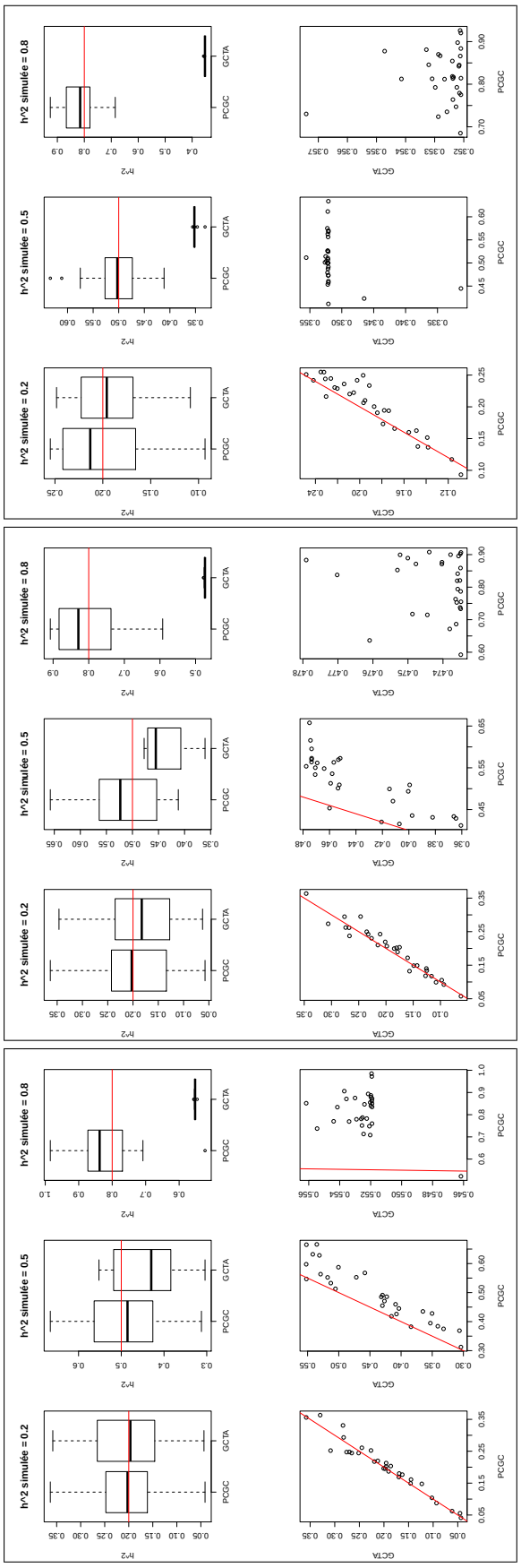
Pour n le nombre d'individus dans l'étude, p le nombre de variants et h^2_{sim} l'héritabilté simulée, les différentes étapes sont :

1. Simulation d'un vecteur de fréquences alléliques $f \sim \mathcal{U}_p(0.05, 0.5)$.
2. Simulation d'un vecteur d'effets des variants $u \sim \mathcal{N}(0_p, \frac{h^2_{sim}}{p}\mathbf{I}_p)$.
3. Tant que $n_{etude} < n$:
   (a) Création d'un vecteur de génotypes d'un individu $Z \sim \mathcal{B}_p(f, 2)$.
   (b) Création du terme génétique de l'individu $g = Zu$.
   (c) Création du terme d'environnement $e \sim \mathcal{N}(0, 1 - h^2_{sim})$
   (d) Création du terme de liability et du phénotype $l = g + e$, $y = \mathbb{1}(l > t)$.
   (e) Sélection dans l'étude : Si $y = 1$ $s \sim \mathcal{B}(P_{cas})$ et si $y = 0$ $s \sim \mathcal{B}(P_{contrôle})$.
   (f) Si $s = 1$, l'individu est ajouté à l'étude et $n_{etude} \leftarrow n_{etude} + 1$. Sinon on reprend à l'étape 3 directement.

6.2.2 Les biais de GCTA dans le cas binaire

Le premier biais montré par Golan et al. [2014] est illustré dans le graphe 6.3. Les auteurs montrent à l'aide de simulations des biais selon les prévalences K et K_e (plus de détail sur les paramètres de simulation dans la légende du graphe). Dans ces graphes sont représentées des estimations d'héritabilité effectuées avec GCTA (qui est le logiciel implantant la méthodologie développée par Lee et al. [2011]) et celles effectués par PCGC. Nous remarquons un biais visible pour GCTA dès le cas $K = 0.01$ et $h^2_{sim} = 0.8$, mais surtout pour les cas $K = 0.001$ et $K = 0.005$ avec $h^2_{sim} = 0.8$. Si le cas $K = 0.005$ et $h^2_{sim} = 0.5$ correspond seulement à une sous-estimation de l'héritabilité, dans les autres cas de sous-estimation, les estimations de GCTA semblent en plus bloquées à une certaine valeur (à environ 0.47 quand $K = 0.005$ et 0.35 quand $K = 0.001$). Ces valeurs correspondent à la valeur de la constante corrective que l'on a décrit en (6.6). En clair dans ces cas, l'héritabilité à l'échelle de l'observation semble égale à 1 (ce qui correspond au paramétrage par défaut de GCTA). Notons que les estimations de PCGC ne sont pas biaisées.

Ces résultats semblent donc montrer que plus K est faible, plus l'héritabilité est sous-estimée par GCTA et que ce phénomène est d'autant plus important que l'héritabilité simulée est importante. Les auteurs décrivent que les biais peuvent également apparaître selon la valeur de K_e. Nous y reviendrons plus tard dans ce chapitre.



(a) K = 1% (b) K = 0.5% (c) K = 0.1%

FIGURE 6.3 – Graphe exposant un biais des méthodes basées sur le REML pour des valeurs de K faibles. Chaque boîte correspond à des simulations avec K fixée à respectivement 1%, 0.5% et 0.01 %. Dans chaque boîte on présente 3 scénarios correspondant à 3 valeurs d'héritabilité simulée (20%, 50% et 80%). Pour chaque scénario, l'étude est de taille n =1000, inclut $p = 10000$ variants causaux et on a $K_e = 0.5$. Pour chaque valeur d'héritabilité, le graphe du haut est un ensemble de boxplots des estimations d'héritabilité avec GCTA et avec PCGC. Le graphe du bas a pour abscisse les estimations d'héritabilité avec PCGC et en ordonnées les estimations de GCTA. La ligne rouge correspond à l'héritabilité simulée sur les graphes du haut et à l'identité pour les graphes du bas.

Une deuxième forme de biais assez étonnante est également décrite par les auteurs. Le graphe 6.4 montre des estimations d'héritabilité avec PCGC (boîte verte) et GCTA (boîte bleue) avec cette fois comme paramètre la taille de l'étude n. Si PCGC donne des résultats satisfaisants et gagnant en précision quand n augmente, ce n'est étonnamment pas le cas de GCTA. Nous observons en effet que les estimations par GCTA décroissent et s'éloignent de plus en plus de la valeur simulée quand n augmente. Ce résultat est très inattendu : si nous nous placions du point de vue de l'apprentissage statistique nous nous attendrions à ce que les résultats soient meilleurs quand on augmente la taille de l'échantillon d'apprentissage mais ce n'est pas le cas ici. Notons que ce biais n'est pas un biais artificiel dû à la forme de nos simulations : la variance des termes génétiques, environnementaux et de liability sont stables quand n augmente (voir annexe D.2).

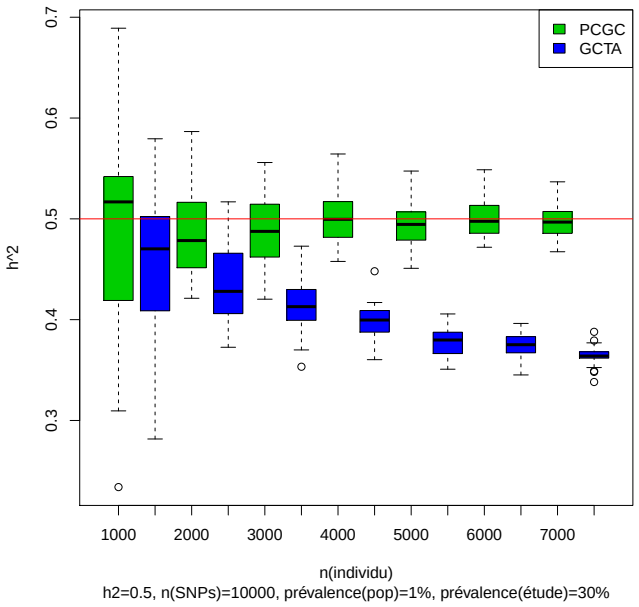


FIGURE 6.4 – Biais des méthodes basées sur le REML quand n augmente. Dans ce graphe l'abscisse représente la taille d'étude $n \in \{1000, 2000, 3000, 4000, 5000, 6000, 7000\}$. Les boîtes bleues correspondent aux estimations par GCTA et les boîtes vertes aux estimations par PCGC. Pour toutes ces études on a $h^2_{sim} = 0.5$ (ligne rouge sur le graphe), $K = 0.01$, $K_e = 0.3$ et $p = 10000$.

6.3 Notre contribution

6.3.1 Discussion des résultats des auteurs de PCGC

Dans leur article [Golan et al., 2014] les auteurs proposent d'attribuer le biais des approches REML quand K et K_e sont très différentes aux interactions GxE présentes

dans la liability à l'échelle de l'étude (voir figure 6.2) : "Just as it was demonstrated .. that the presence of GxG interactions leads to underestimation of the fraction of heritability explained, we suspected that the presence of such induced GxE interactions might result in underestimation of the heritability" et "The magnitude of the bias increases when (i) the disease is rarer, (ii) the proportion of cases is closer to half, and (iii) the heritability is higher. Indeed, these circumstances each increase the induced GxE interaction." Par ailleurs les auteurs utilisent de manière équivalente les termes de corrélation GxE et d'interaction GxE, ce qui nous a surpris. Dans cette sous-section nous avons tenté de vérifier ces affirmations en passant par des simulations.

Nous allons regarder l'influence des paramètres n, K_e et h^2_{sim} (notons donc qu'ici K sera une quantité fixe) sur la variance du terme génétique, du terme environnemental et de la liability ainsi que sur la corrélation GxE. Pour chacune des combinaisons de $K_e \in \{0.3, 0.5, 0.8\}$, $h^2_{sim} \in \{0.3, 0.5, 0.8\}$ et $n \in \{500, 2000, 3000, 4000, 5500\}$ nous avons simulé 20 études suivant l'approche décrite en 6.2.1 avec $K = 0.05$ et $p = 5000$.

Tous les résultats de cette sous-section se présenteront sous la forme de figure décomposable en grille d'ensemble de boxplots : les lignes et colonnes correspondent respectivement aux différentes valeurs d'h^2_{sim} et de K_e. Chacune de ces sous-figures est un ensemble de boxplots et l'axe des abscisses correspond aux différentes valeurs de n.

Nous avons représenté dans la figure 6.5 l'influence des 3 paramètres sur l'estimation d'héritabilité. Nous y retrouvons des résultats déjà connus à savoir que les biais de GCTA semblent augmenter avec n et l'héritabilité, mais également que ces biais sont une fonction non linéaire de K_e. Ce graphe corrobore les résultats de Golan et al. qui avaient observé que le biais était maximal quand $K_e = 0.5$.

Enfin dans la figure 6.6, nous avons regardé l'influence des paramètres sur la corrélation GxE. La première chose que nous remarquons est qu'encore une fois n n'a pas d'influence. Nous observons également que pour $K_e = 0.8$ la corrélation est presque nulle. De plus à h^2_{sim} fixée, la corrélation est similaire entre $K_e = 0.3$ et $K_e = 0.5$, bien que légèrement plus élevée pour $K_e = 0.5$. À K_e fixée, elle est maximale pour $h^2_{sim} = 0.5$ et minimale pour $h^2_{sim} = 0.8$. Ces résultats peuvent s'expliquer par le fait que la corrélation GxE, créée par l'enrichissement en cas dans l'étude par rapport à la population générale et illustrée par l'écart entre le groupe de cas et celui des contrôles sur la figure 6.2, est d'autant plus forte lorsque les deux groupes de l'étude et l'importance des deux composantes G et E sont équilibrés.

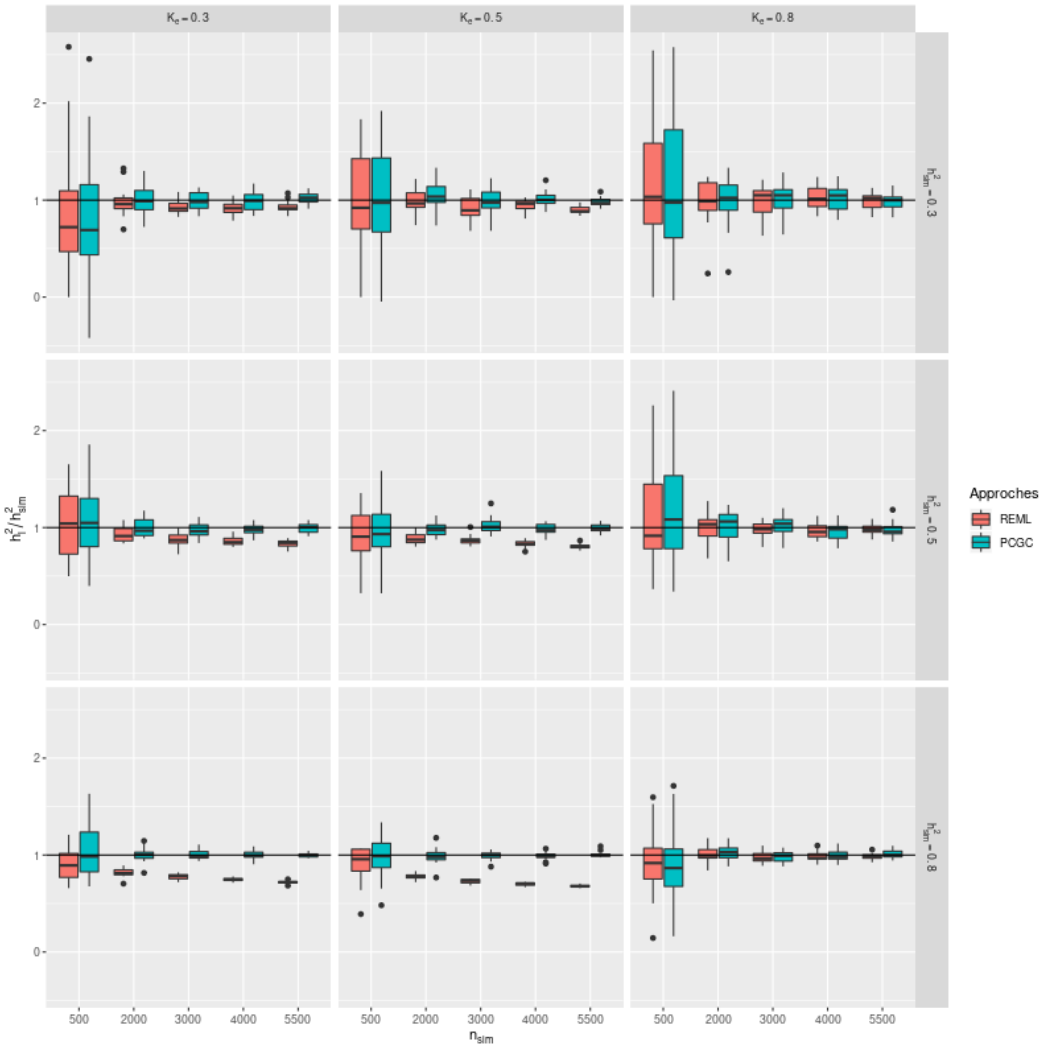


FIGURE 6.5 – Graphe d'influence de n, h^2_{sim} et K_e sur l'estimation d'héritabilité pour le cas binaire. Les lignes et colonnes de la grille de boxplots correspondent respectivement aux différentes valeurs d'h^2_{sim} et de K_e. Dans chacun des ensembles de boxplots, l'axe des abscisses correspond aux différentes valeurs de n et l'axe des ordonnées au rapport h^2_{est}/h^2_{sim}. Les boîtes bleues correspondent aux estimations par PCGC et les boîtes rouges aux estimations par GCTA.

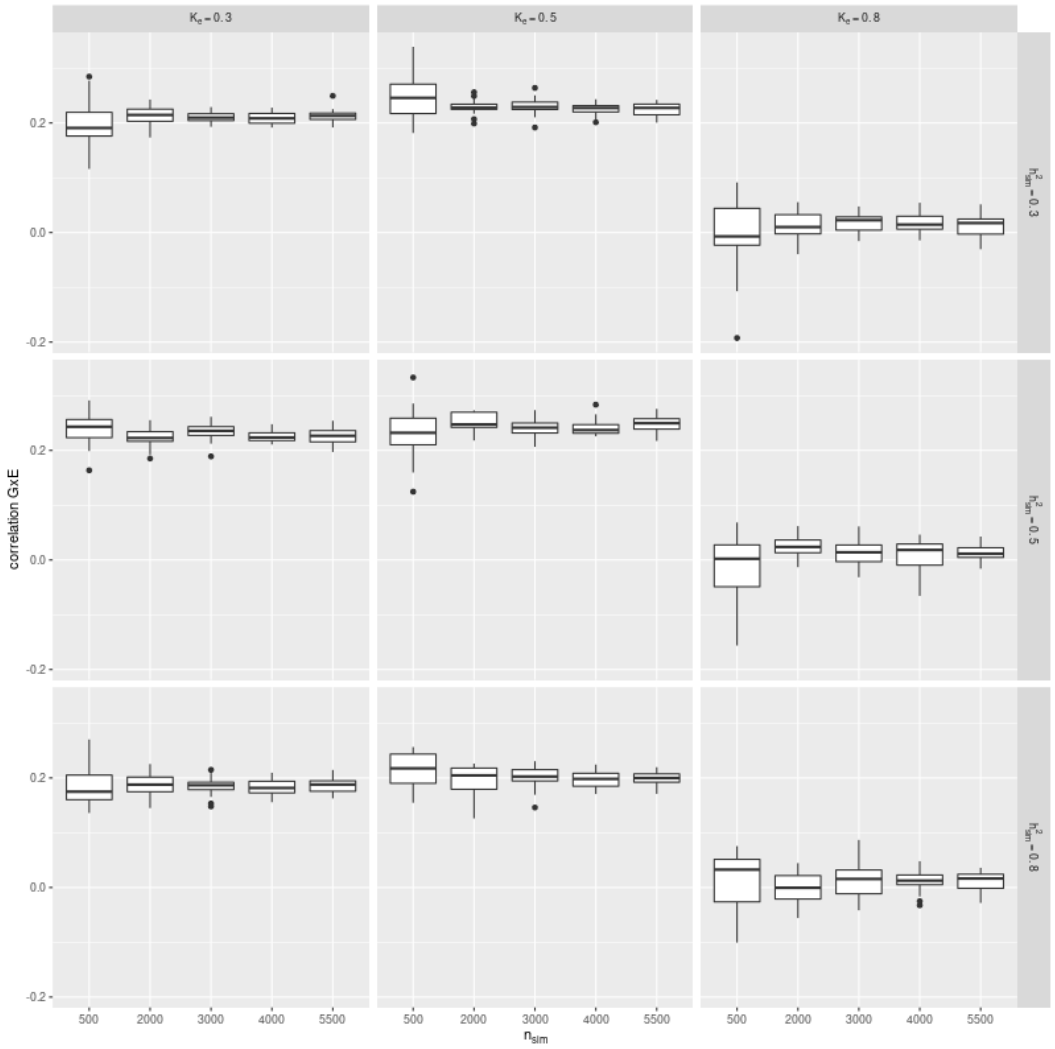


FIGURE 6.6 – Graphe d'influence de n, h^2_{sim} et K_e sur la corrélation GxE. Les lignes et colonnes de la grille d'ensemble de boxplots correspondent respectivement aux différentes valeurs d'h^2_{sim} et de K_e. Dans chacun des boxplots, l'axe des abscisses correspond aux différentes valeurs de n et l'axe des ordonnées à la corrélation GxE.

Au vu de nos résultats il semble bien y avoir un lien entre biais de GCTA et présence de corrélation GxE. En particulier quand $K_e = 0.8$, il n'y a ni corrélation GxE, ni biais sur l'héritabilité. À l'inverse lorsque $K_e = 0.5$, le biais et la corrélation sont maximaux comme cela avait été observé par Golan et al. [2014]. Toutefois il nous semble trop hâtif de conclure que cette corrélation soit la seule explication. Nous ignorons la nature profonde du lien entre corrélation GxE et biais de l'héritabilité mais nous observons qu'il n'est pas linéaire : à $K_e = 0.5$ et $K_e = 0.3$ le biais de l'héritabilité augmente avec h^2_{sim} et ce n'est pas le cas de la corrélation qui est maximale pour $h^2 = 0.5$, contrairement à ce que suggéraient Golan et al. [2014]. Le lien entre une héritabilité élevée et un biais d'estimation plus important pourrait s'expliquer, quant à lui, par le fait que l'estimation de l'héritabilité est bornée à 1 à l'échelle de l'observation.

Ces résultats liant bais et corrélation GxE peuvent sembler surprenants car cette corrélation mène probablement plutôt à une sur-estimation de l'héritabilité à l'échelle de la liability, du fait de la confusion entre environnement partagé et génétique partagée. Cependant, la corrélation est associée à une interaction GxE artificielle sur le phénotype à l'échelle de l'observation (i.e. après seuillage de la liability), étant donné que les cas de l'étude ont à la fois un terme génétique et un terme environnemental élevés. Cela explique sans doute la confusion entre corrélation et interaction dans l'article de Golan et al. [2014]. Cette interaction artificielle pourrait en effet expliquer la sous-estimation de l'héritabilité à l'échelle de l'observation. Notons que ce raisonnement n'explique pas du tout le biais de GCTA observé par Golan et al. quand la taille de l'ensemble d'apprentissage augmente.

6.3.2 Une application aux données de cardiomyopathie

Nous avons appliqué PCGC et GCTA sur des données de cardiomyopathie dilatée pour une évaluation des méthodes sur données réelles.

Estimation d'héritabilité avec plusieurs prétraitements

Ici nous sommes partis des estimations d'héritabilité sur la cardiomyopathie dilatée obtenues avec GCTA par Mathilde Carlier pendant son stage. Nous avons décidé de reprendre son travail pour y ajouter l'estimation d'héritabilité par PCGC. Comme nous allons le voir nous avons obtenu des résultats problématiques avec PCGC tels que des héritabilités de 300 %. Après enquête le contrôle qualité avant estimation d'héritabilité joue un rôle très important en binaire et nous présentons ici nos résultats.

La cardiomyopathie dilatée est une maladie assez rare ($K \simeq 1/3000$) [Garnier et al., 2020] qui se traduit par une augmentation de la taille des cavités cardiaques alors que les parois musculaires restent minces (dilatation des ventricules) et par une faiblesse de la contraction du muscle cardiaque (hypokinésie des ventricules) qui peut entraîner une réduction du débit sanguin. Cette maladie existe sous une forme monogénique (environ 30 % des cas) et une forme complexe qui va nous intéresser ici.

Nous disposons de données de GWAS sur la cardiomyopathie dilatée. Ces données de puce contiennent les génotypes d'environ 700 000 variants (génotypage réalisé avec la puce Illumina OmniExpress) pour 3084 individus répartis en 3 populations (1790 français, 1120 allemands et 174 italiens). Au total l'étude contient environ 2/3 de cas et 1/3 de témoins. Notons que nous sommes conscients que ce type de mélange de population est normalement à éviter car pouvant entraîner des biais, mais nous avons choisi de passer outre pour avoir une taille d'échantillon la plus grande possible. Il aurait été toutefois intéressant d'effectuer cette analyse uniquement sur les Français, puisque cette population présente un effectif "suffisant" (contrairement aux Italiens) et une répartition de cas et de témoins (contrairement à la population allemande qui n'est composée que de cas).

Nous commencerons par décrire les filtres qu'avait utilisés Mathilde Carlier. Ce filtrage a été réalisé avec PLINK 1.9 [Purcell, 2009].

— Suppression des individus avec un callrate < 0.9 (0 individu supprimé).

— Suppression des variants avec un callrate < 0.9 (1173 variants supprimés).

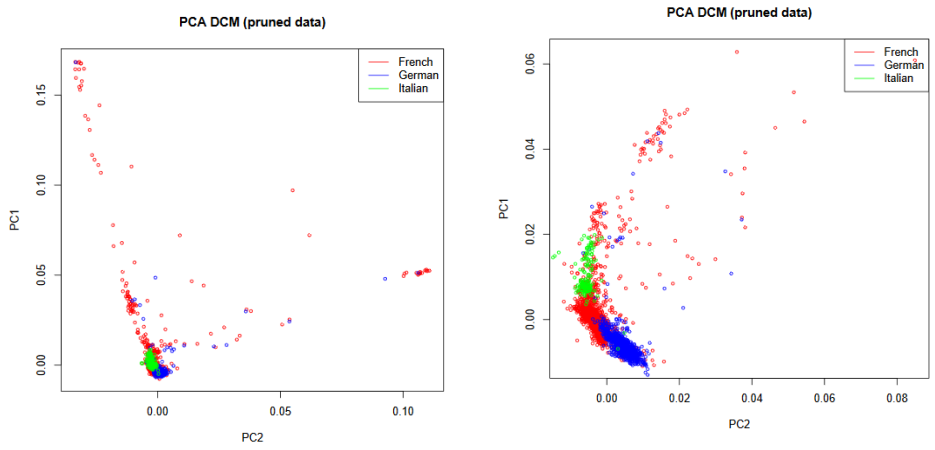
— Suppression des variants avec une p-value au test de Hardy-Weinberg sur les contrôles ≤ 0.001 (13719 variants supprimés).

— Suppression des individus dont la valeur IBD est ≥ 0.185 pour travailler sur des individus non-apparentés (24 individus supprimés).

— Détermination de 155803 SNPs en équilibre de liaison (avec une fenêtre glissante de 50 SNPs, un pas de 5 SNPs et un seuil de R^2 fixé à 0.2).

— Calcul d'une analyse en composantes principales sur la matrice de ressemblance génétique des individus avec les variants en équilibre de liaison. Exclusion des individus outliers selon les deux premières composantes principales (43 individus supprimés) puis nouveau calcul de l'ACP. Les deux première composantes principales de chaque ACP sont représentées dans le graphe 6.7.

Après ces filtres, il nous reste 3006 individus et 155803 variants.



(a) Les deux premières composantes principales avant suppression des individus aberrant

(b) Les deux premières composantes principales après suppression des individus aberrant

FIGURE 6.7 – Graphe des deux premières composantes principales pour les données de cardiomyopathie pour le premier jeu de filtres. Le graphe de gauche correspond à l'ACP avant l'exclusion des individus outliers et le graphe de gauche correspond à l'ACP après exclusion. Ces composantes principales sont calculées sur les variants en équilibre de liaison.

Au vu du graphe 6.7 nous voyons qu'il existe une certaine structure de population sous-jacente. Pour la prendre en compte les 3 premières composantes principales sont ajoutées à l'estimation d'héritabilité. Nous estimons également l'héritabilité sans prendre en compte cette structure pour voir la robustesse de l'estimation. Un point que nous devons noter est que dans le travail de Mathilde Carlier l'analyse en composantes principales a été réalisée sur les variants en équilibre et l'estimation d'héritabilité également pour une raison bioinformatique : permettre à des algorithmes de machine learning de tourner en des temps raisonnables sur le cluster de calcul du CNRGH (résultats non présentés). Il aurait été plus rigoureux de calculer l'héritabilité sur l'ensemble des variants. Cependant pour avoir des valeurs comparables nous allons continuer nos analyses sur l'ensemble des variants en équilibre de liaison, mais nous avons conscience qu'elles ont un intérêt biologique limité.

Les résultats sont résumés dans la ligne "Jeu de filtre 1" de la table 6.1. Nous voyons que pour les méthodes basées sur le REML nous obtenons une héritabilité d'environ 0.32 avec ou sans composantes principales intégrées dans le modèle. Pour PCGC les résultats sont très étonnants : avec ou sans composantes principales les résultats sont très au dessus de 1, ce qui n'a aucun sens biologique. Notons également que les écarts-types (obtenus avec une approche de ré-échantillonnage dite jackknife) sont très élevés. Nous voyons clairement qu'il y a un problème pour PCGC puisque les héritabilités au dessus de 1 n'ont aucun sens. Mais nous observons également un problème pour les

approches REML : en effet la valeur d'héritabilité obtenue correspond à une héritabilité à l'échelle de l'observation estimée à 1. Cela révèle le plus grand problème de l'approche de Lee et al. [2011] : leur estimation d'héritabilité est (par défaut, nous en reparlerons ensuite) bornée par la constante corrective.

Jeu de filtre	**Méthodes**	**Sans Composante Principales**	**Avec 3 Composantes Principales**
Jeu de filtre 1	REML	0.32(0.03)	0.32(0.03)
	PCGC	**3.01(0.28)**	**5.51(0.93)**
Jeu de filtre 1 + Jeu de filtre 2	REML	0.33(0.03)	0.18(0.04)
	PCGC	**5.81(0.34)**	0.17(0.07)

Tableau 6.1 – Table d'estimation d'héritabilité sur les données de cardiomyopathie.

Les résultats ne sont pas du tout satisfaisants : les valeurs de PCGC sont complètement aberrantes, et celles de GCTA ne sont pas satisfaisantes non plus (car elles correspondent à la valeur maximale de l'algorithme et donc il est difficile de savoir si l'algorithme a convergé). Nous avons trouvé dans la littérature que des filtres plus stricts étaient conseillés pour l'estimation d'héritabilité de maladies complexes [Lee et al., 2011]. Désormais nous nous référerons aux premiers filtres avec la dénomination "jeu de filtres 1" et à ceux (plus stricts) que nous allons présenter maintenant avec la dénomination "jeu de filtres 2" (qui s'ajoutent au "jeu de filtres 1").

Le jeu de filtres 2 est le suivant :

— Suppression des individus avec un callrate $\geq 99\%$ (75 individus supprimés).

— Suppression des variants avec un callrate $\geq 99\%$ (7 264 variants supprimés).

— Supression des variants avec une MAF $\leq 5\%$ (52 648 variants supprimés).

— Suppresion d'un individu parmi les paires d'individus trop apparentés au sens de la GRM (GRM ≥ 0.025, 204 individus supprimés). Les auteurs de GCTA [Yang et al., 2010] insistent sur l'importance de ce filtre.

Après les deux sessions de filtres il nous reste 2 740 individus et 90 092 SNPs. Nous calculons encore une fois les composantes principales pour rechercher une structure de population sous-jacente. Les deux premières composantes sont représentées dans le graphe 6.8. Les graphes sont plutôt regroupés mais nous voyons qu'il y a une très claire structure de population. Nous ajouterons donc 3 composantes principales pour l'estimation de l'héritabilité.

Les résultats avec les deux jeux de filtres sont présentés dans la deuxième ligne de la table 6.1. Nous y voyons que les estimations d'héritabilité ne sont pas acceptables

si nous n'ajoutons pas de composantes principales. A l'inverse avec les composantes principales, les estimations sont plutôt satisfaisantes : les deux méthodes donnent une valeur similaire (0.18 pour les approches REML et 0.17 pour PCCG) et cette valeur n'est pas aberrante.

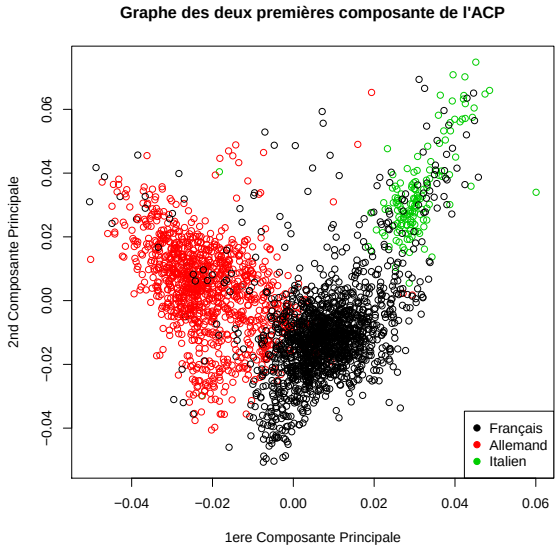


FIGURE 6.8 – Graphe des deux premières composantes principales pour les données de cardiomyopathie dilatée pour les deux jeux de filtres cumulés.

Influence des différents filtres

L'ajout du deuxième jeu de filtres a permis d'obtenir des valeurs d'héritabilité acceptables. Ces résultats montrent l'importance du prétraitements des données avant l'estimation d'héritabilité dans le cas binaire [Lee et al., 2011]. Nous avons ensuite décidé d'enquêter plus en détail sur l'influence de chacun des filtres du deuxième jeu. Les résultats sont compilés dans la table 6.2. La première et la 5ème ligne correspondent aux résultats de la table 6.1.

La deuxième ligne ajoute un filtre sur la MAF aux premiers filtres. Sans l'ajout de composantes principales l'estimation par PCGC est au dessus de 1 et n'est donc pas acceptable. Pour GCTA l'héritabilité est estimée par la valeur de la constante et nous ne pouvons donc pas conclure si la valeur est pertinente. En ajoutant les CP, l'estimation par PCGC devient acceptable à 0.31(0.06) et celle des approches REML est à environ 0.22(0.04). Les deux valeurs sont différentes mais les valeurs sont "acceptables".

La troisième ligne correspond au jeu de filtres 1 + les filtres sur le callrate (individus et variants). Nous voyons que pour PCGC les valeurs d'héritabilité sont très au dessus

de 1 avec ou sans CP. Pour les approches REML, l'estimation correspond à la valeur maximale et il est donc compliqué de conclure. Nous en déduisons que ce filtre n'est sans doute pas le plus important.

Enfin la quatrième ligne montre l'influence du filtre avec seuil sur l'apparentement. Pour les méthodes REML le filtre ne semble rien apporter, les estimations d'héritabilité restent au niveau de la constante d'ajustement. Pour les estimations avec PCGC sans les CP la valeur n'est pas acceptable mais elle le devient si on ajoute les CP avec une estimation de l'héritabilité à environ 0.31(0.11).

Une synthèse de nos résultats est que les filtres les plus importants pour obtenir des valeurs acceptables pour PCGC sont le filtre MAF et le filtre sur l'apparentement. Pour l'approche REML le filtre le plus important semble être celui sur la MAF, même si nous ne pouvons pas vraiment conclure pour les autres filtres du fait de la limitation des estimations d'héritabilité par la valeur de la constante. Le filtre sur le callrate semble peu jouer dans les deux approches. Nous remarquons également que même après les filtres, les deux approches restent sensibles à une structure de population sous-jacente : PCGC en particulier ne rend aucune valeur d'héritabilité acceptable si nous n'incluons pas les CP dans l'estimation d'héritabilité (pour les approches REML il n'est pas possible d'avoir des conclusions arrêtées).

Jeu de filtre	**Méthodes**	**Sans Composantes Principales**	**Avec 3 Composantes Principales**
Jeu de filtre 1	REML	0.32(0.03)	0.32(0.03)
	PCGC	3.01(0.28)	5.51(0.93)
Jeu de filtre 1 + filtre MAF	REML	0.32(0.03)	0.22(0.04)
	PCGC	6.40(0.46)	0.31(0.06)
Jeu de filtre 1 + filtre callrate	REML	0.32(0.03)	0.32(0.03)
	PCGC	3.09(0.22)	5.89(1.71)
Jeu de filtre 1 + filtre GRM	REML	0.33(0.03)	0.33(0.04)
	PCGC	8.57(0.44)	0.31(0.11)
Jeu de filtre 1 + Jeu de filtre 2	REML	0.33(0.03)	0.18(0.04)
	PCGC	5.81(0.34)	0.17(0.07)

Tableau 6.2 – Table sur l'influence des différents filtres pour l'estimation d'héritabilité sur les données de cardiomyopathie dilatée.

Influence du nombre de composantes principales

Nous remarquons que nous avons mis 3 composantes principales comme covariables fixes dans le modèle. Il n'y a pas de règles absolues pour le choix du nombre de com-

posantes à ajouter. Nous avons donc décidé de regarder l'influence du nombre de composantes ajoutées sur l'estimation de l'héritabilité dans le cas binaire. Ces travaux font écho à ceux de Dandine-Roulland [2014] qui se sont intéressés au cas continu.

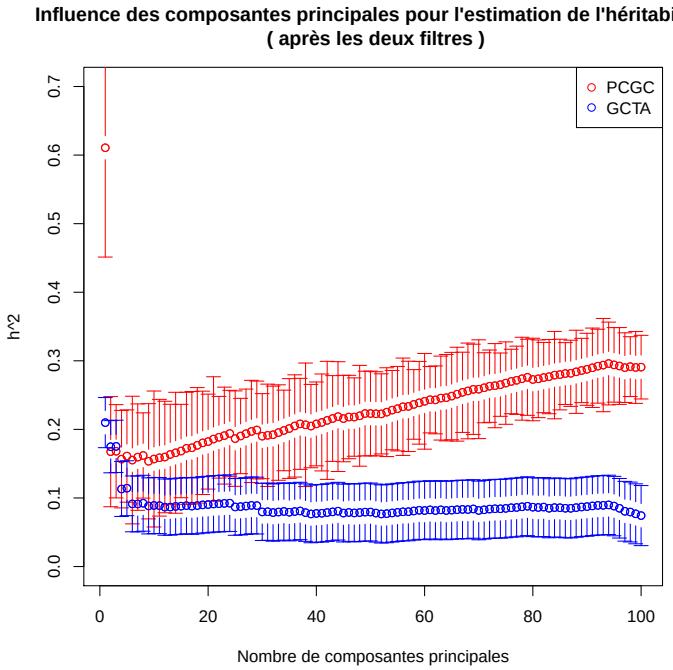


FIGURE 6.9 – Graphe sur l'influence des composantes principales pour l'estimation d'h^2 en binaire. L'abscisse représente le nombre de composantes principales et l'ordonnée les estimations d'héritabilité sur les données de cardiomyopathie. Le rouge représente les estimations de PCGC et le bleu les estimations de GCTA. Pour chaque nombre de CP on représente un point par estimation d'héritabilité. Les barres d'erreur correspondent à un écart type rendu par les méthodes.

Les résultats sont représentés dans le graphe 6.9. Commençons par les résultats pour les approches REML (courbe bleue) : nous distinguons une variance dans les estimations quand le nombre de CP est inférieur à 5. Ensuite l'estimation semble se stabiliser à environ 0.1 et nous distinguons une légère décroissance à partir d'environ 95 composantes principales. Enfin notons que les intervalles de confiance sont stables (ce qui n'est pas très étonnant, les écarts-types calculés par GCTA sont proportionnels à 1/n [Visscher and Goddard, 2015]).

Pour les estimations de PCGC nous remarquons qu'à l'exception des estimations incluant moins de 4 CP, il y a augmentation des estimations d'héritabilité d'environ 0.18 à 0.25 quand le nombre de CP augmente. Sur la toute fin de la courbe on remarque une stabilisation (voir même une légère décroissance). Nous remarquons également que les écarts-types (estimés par une approche jackknife) diminuent avec le nombre de composantes ajoutées.

Au vu des résultats pour les approches REML il semble raisonnable de prendre entre 5 et 10 composantes principales pour avoir des estimations stables. Par contre pour PCGC il est difficile de conclure puisque la courbe est croissante. Un point noir de ces résultats est que l'on utilise les mêmes données pour estimer les composantes principales et pour estimer l'héritabilité. Il est donc difficile de différencier suppression de structure de population et suppression d'information génétique pertinente pour le trait étudié.

6.3.3 Une explication pour les valeurs aberrantes de PCGC

Dans cette section nous donnerons une explication graphique de PCGC qui permettra d'expliquer simultanément les valeurs aberrantes obtenues sur données réelles et l'importance du filtre sur les coefficients de la GRM.

Nous avons vu plus haut que la méthode PCGC consiste en une régression linéaire entre les produits de phénotypes normalisés et la ressemblance génétique entre individus puis en une multiplication du coefficient de régression par une correction. Il est donc très facile d'avoir une représentation "graphique" de PCGC. Dans la suite de cette section nous allons travailler sur une étude simulée avec comme paramètres de simulation $n = 150$, $K = 0.05$, $K_e = 0.62$, $p = 10000$ et $h^2_{sim} = 0.7$. Pour cette étude, l'héritabilité estimée par PCGC vaut 0.71. La représentation graphique de PCGC "normal" est la figure 6.10c. Sur cette figure l'axe des abscisses correspond aux coefficients de GRM et l'axe des ordonnées au produit de phénotypes normalisés : les points d'ordonnée -1 correspondent aux paires cas/contrôle, les points d'ordonnée 0.6 correspondent aux paires de contrôles et ceux d'ordonnée 1.5 aux paires de cas.

Nous allons ajouter un coefficient δ à tous les coefficients de la GRM pour les paires de cas. Cela revient à jouer sur l'apparentement des cas (les cas seront plus ou moins apparentés selon le signe de δ). Nous sommes bien conscients que cela a peu de sens biologique car nous supposons uniquement une augmentation de l'apparentement pour les cas mais nous l'avons fait pour faire apparaître de manière simple les instabilités de PCGC.

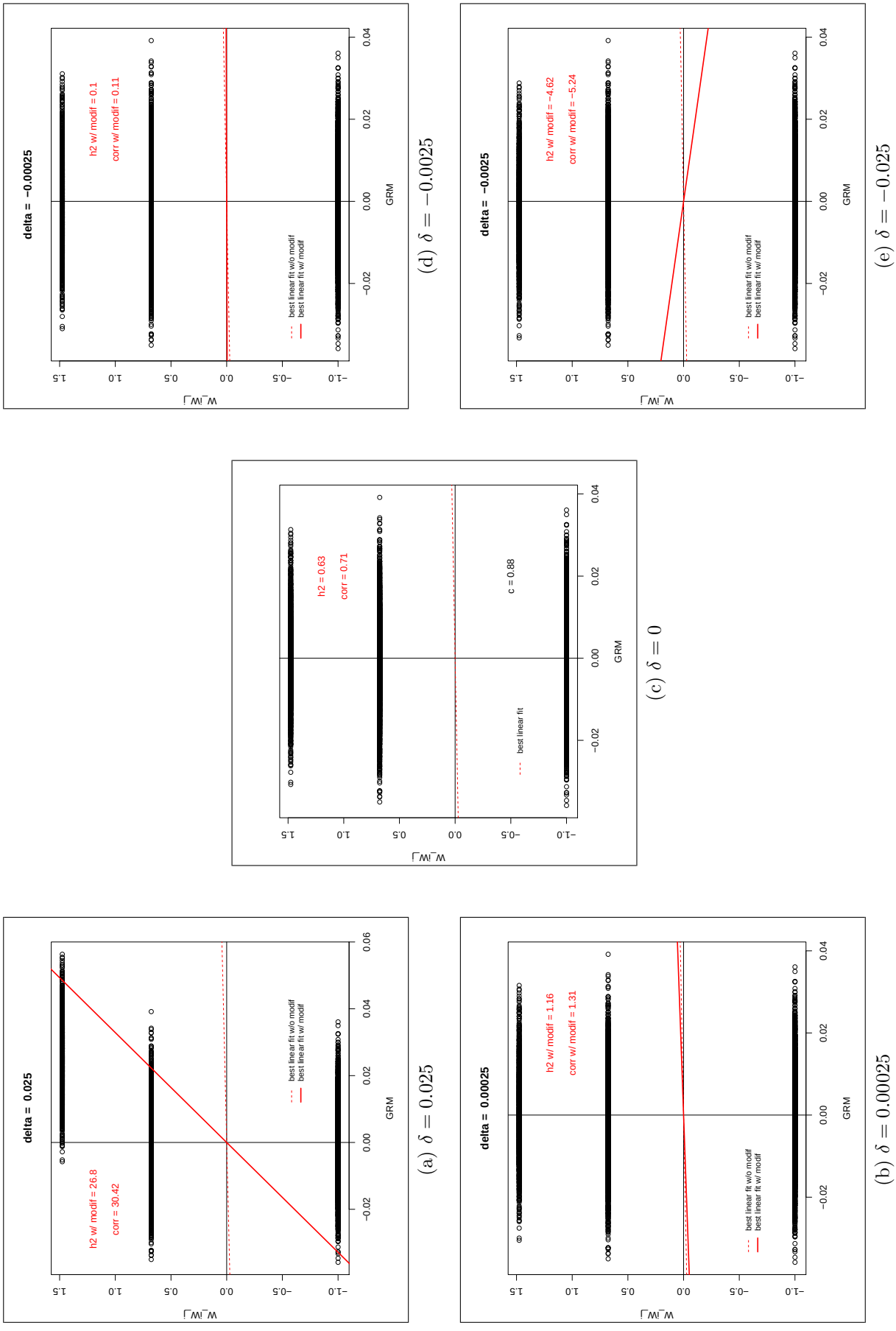
Le graphe 6.10c correspond à la régression de PCGC quand $\delta = 0$. Nous obtenons alors une héritabilité de 0.63 qui va nous servir de référence. Dans le graphe 6.10b nous ajoutons un $\delta = 0.00025$ sur les paires de cas. Nous avons choisi cette valeur car elle correspond à un centième du seuil empirique proposé dans [Yang et al., 2010] pour filtrer

les individus trop apparentés. En pratique, nous "tirons vers la droite" les coefficients de la GRM des paires de cas. L'estimation d'héritabilité par PCGC après l'ajout de cette petite quantité est de 1,16. En augmentant d'une toute petite quantité la ressemblance génétique entre cas nous avons quasiment multiplié par deux l'estimation d'héritabilité. PCGC semble donc très sensible aux apparentements cryptiques. Dans la figure 6.10a, δ est fixé à 0.025 ce qui est une valeur énorme mais que nous allons utiliser juste pour montrer la sensibilité à δ. Pour ce scénario nous estimons l'héritabilité à 26,8 : il est donc facile d'atteindre des valeurs complètement aberrantes d'héritabilité.

Les $\delta < 0$ ont également une influence. La figure 6.10e correspond à $\delta = -0.0025$ et donne une estimation d'héritabilité de 0.1. Ainsi un manque de ressemblance génétique entre cas semble entraîner un biais négatif pour sur l'estimation d'héritabilité. Le cas 6.10e correspond au cas $\delta = -0.025$ et donne une estimation d'héritabilité négative de -4.62, montrant que nous pouvons également avoir des valeurs aberrantes négatives.

Ces ajouts de δ semblent donner une bonne intuition des biais que peut avoir PCGC et de son extrême sensibilité à une structure de population cryptique. Nous avons également testé cette approche en ajoutant un δ sur les paires de contrôles et sur les paires cas / contrôles et nous avons obtenus des résultats cohérants. Ces comportements sont résumés dans la table 6.3. Si $\delta > 0$ pour les paires de phénotypes concordants, alors l'estimation de l'héritabilité augmente. C'est assez cohérent : on pratique une régression pour faire apparaître le lien entre ressemblances génétique et phénotypique. Si les individus avec un même phénotype ont une ressemblance génétique forte alors le lien augmente et il est donc logique qu'une ressemblance génétique augmentée entre phénotypes identiques augmente ce lien et donc l'héritabilité (et qu'à l'inverse une ressemblance réduite avec $\delta < 0$ diminue le lien). Au contraire pour les paires de phénotypes discordants, une forte ressemblance génétique pousse à diminuer le lien entre ressemblances génétique et phénotypique.

Notons enfin que nous observons les mêmes comportements pour les estimations avec les approches REML, ce qui est cohérent car les deux approches se ressemblent mais nous n'avons pas d'explication aussi intuitive pour les approches REML.



(a) $\delta = 0.025$ (b) $\delta = 0.00025$ (c) $\delta = 0$ (d) $\delta = -0.0025$ (e) $\delta = -0.025$

FIGURE 6.10 – Ensemble de graphes pour illustrer les valeurs aberrantes de PCGC. Chacun des sous-graphes (a), (b), (c), (d) et (e) l'axe des abscisses correspond aux coefficients de GRM et l'axe des ordonnées au produit de phénotypes normalisés avec un delta de respectivement 0.025, 0.0025, 0, -0.0025 et -0.025 pour les paires de cas. Les droites pleines et en pointillés correspondent à la régression linéaire entre les coefficients de GRM et les produits de phénotype normalisés avec et sans application du delta. Sur chacun des sous-graphe nous avons également écrit l'estimation d'héritabilité après application du delta et la valeur du coefficient de la régression linéaire.

	2 phénotypes concordants	2 phénotypes discordants
$\delta > 0$	$h^2 \nearrow$	$h^2 \searrow$
	$(h^2 \underset{\delta \to +\infty}{\to} 0+)$	$(h^2 \underset{\delta \to -\infty}{\to} 0-)$
$\delta < 0$	$h^2 \searrow$	$h^2 \nearrow$
	$(h^2 \underset{\delta \to -\infty}{\to} 0-)$	$(h^2 \underset{\delta \to +\infty}{\to} 0+)$

TABLEAU 6.3 – Table d'influence de δ sur les estimations de PCGC

6.4 En résumé ...

Dans cette section nous avons proposé une discussion autour d'un biais des méthodes d'estimation d'héritabilité pour le cas binaire basées sur les méthodes REML. Nous sommes en particulier revenus sur l'explication proposée par Golan et al. de ce biais avec une corrélation ou interaction GxE qui nous semblait confuse. Nous avons appuyé nos idées par un travail sur des simulations.

Nous avons également montré la grande sensibilité de l'estimation d'héritabilité aux filtres de prétraitement des données avec une application sur des données de GWAS pour la cardiomyopathie dilatée. Les estimations d'héritabilité étaient aberrantes en l'absence de filtres stricts, et acceptables avec ces filtres stricts. Nous avons également discuté du nombre de composantes principales à inclure comme covariables de structure dans le modèle, et concluons qu'il semble judicieux d'en prendre entre 5 et 10.

Enfin nous avons proposé une explication pour les valeurs aberrantes de la méthode PCGC, en observant le comportement de l'estimation d'héritabilité après avoir légèrement augmenté la ressemblance génétique (au sens de la GRM) des paires de cas.

Chapitre 7

Perspectives

Dans ce chapitre nous présenterons rapidement une conclusion générale de ce travail, puis des possibles continuations. Nous proposerons une méthode pour obtenir une distribution de paramètre de pénalisation, puis discuterons de l'utilisation de la régression ridge hétéroscédastique pour permettre la modélisation de plusieurs effets génétiques, mais également pour permettre l'inclusion de covariables non-pénalisées dans la régression ridge. Nous présenterons ensuite nos idées pour une vérification de l'approximation que nous avons proposée au chapitre 5 sur une plus grande plage de ratio n/p. Ensuite nous discuterons très brièvement de comment quitter le modèle linéaire, et terminerons par une proposition pour essayer de corriger une partie des biais des approches basées sur le REML pour le cas binaire.

7.1 Un résumé de la thèse en quelques lignes

Dans cette thèse, nous nous sommes intéressés à l'utilisation de la régression ridge sur des données génétiques et phénotypiques de population humaine. L'objectif initial de cette thèse était de lier, grâce à une méthode d'apprentissage statistique, les concepts d'héritabilité et de prédiction d'un phénotype, deux concepts qui semblent liés mais définis dans des mondes différents. Les méthodes d'estimation d'héritabilité modernes se basant sur les modèles à effets aléatoires, nous avons décidé d'utiliser comme méthode d'apprentissage la régression ridge car elle permet des liens directs avec ces modèles.

Après avoir présenté dans les deux premiers chapitres l'apprentissage statistique et la régression ridge, nous avons présenté dans le chapitre 3 nos solutions à des problèmes calculatoires spécifiques au cadre de la grande dimension.

Nous avons alors proposé plusieurs méthodes d'estimation d'héritabilité toutes dérivées de la régression ridge. Parmi elles, une méthode basée sur une transformation de

l'hyperparamètre optimal de la régression ridge a des résultats équivalents aux estimations basées sur les modèles à effets aléatoires. Ce résultat s'est vérifié à la fois sur des données simulées et réelles.

Nous nous sommes ensuite intéressés au comportement de plusieurs mesures de la capacité de prédiction de la régression ridge en fonction du ratio des dimensions de l'ensemble d'apprentissage et de l'héritabilité. Pour chacune d'entre elles, nous avons proposé une approximation selon l'héritabilité et le ratio des dimensions de l'ensemble d'apprentissage, et vérifié la validité de ces approximations sur des données simulées et réelles.

En conclusion, nous avons montré que la régression ridge permettait des estimations de l'héritabilité tout à fait satisfaisantes pour un contexte de GWAS chez l'homme et que la capacité de prédiction de la régression ridge était une fonction croissante de l'héritabilité. En revanche, nous avons montré que dans le contexte de grande dimension des GWAS, la régression ridge ne donnait pas de bonnes prédictions même si l'héritabilité du trait à prédire était élevée.

7.2 Distribution de paramètre de pénalisation optimaux pour la régression ridge

Nous allons présenter ici une approche permettant d'obtenir à bas coût une distribution du paramètre de pénalisation optimal pour la régression ridge. Nous avions montré dans la section 2.6.1 que le paramètre de pénalisation optimal est théoriquement le même entre le modèle original et un modèle contrasté après multiplication par une matrice vérifiant les propriétés de contraste. Si nous disposions d'un ensemble de telles matrices, nous pourrions alors en déduire un ensemble d'estimations du paramètre de pénalisation optimal. Dans cette section, nous utiliserons et calculerons des matrices de contrastes (malgré l'absence d'effets fixes) pour obtenir une distribution de paramètre de pénalisation optimal.

Supposons que nos données suivent le modèle linéaire sans intercept ou effet fixes

$$\mathbf{y} = \mathbf{Z}u + \mathbf{e} \tag{7.1}$$

avec $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ la matrice de données standardisée, $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ sa décomposition en valeurs singulières, $u \in \mathbb{R}^p$ et $\mathbf{e} \in \mathbb{R}^n$. Pour simplifier nous supposerons que $\mathbf{Z}$ n'est pas standardisé empiriquement et en conséquence ces valeurs singulières sont non nulles.

Notre objectif est d'estimer le paramètre de pénalisation optimal avec la GCV. Soit $\mathcal{P}_{N_c}(\llbracket 1, n \rrbracket)$ l'ensemble des N_c-combinaisons de $\llbracket 1, n \rrbracket$ et $\mathcal{E} = \{\mathcal{E}_1, ..., \mathcal{E}_{N_c}\} \in \mathcal{P}_{N_c}(\llbracket 1, n \rrbracket)$ un de ces élément. Posons $\mathbf{C}_\mathcal{E} = \left[\mathbf{U}_{\mathcal{E}_1}, ..., \mathbf{U}_{\mathcal{E}_{N_c}}\right]^T \in \mathcal{M}_{N_c,n}$ une matrice définie comme la concaténation en lignes d'une combinaison de N_c colonnes de $\mathbf{U}$. Cette matrice vérifie $\mathbf{C}_\mathcal{E}\mathbf{C}_\mathcal{E}^T = \mathbf{I}_{N_c}$ et $\mathbf{C}_\mathcal{E}\mathbf{0}_n = \mathbf{0}_{N_c}$ (puisque l'on suppose que le modèle est sans effets fixes) et donc les propriétés de contraste du modèle. Nous avons vu dans la section 3.3.1 que les calculs de la GCV se simplifiaient dans un modèle contrasté par une combinaison de vecteurs colonne de la matrice $\mathbf{U}$. En résumé il est techniquement possible d'obtenir une estimation du paramètre de pénalisation optimal à partir de n'importe quelle combinaison de vecteurs colonne de $\mathbf{U}$.

Avoir une distribution de ce paramètre pourrait avoir de nombreuses utilités. La plus évidente serait d'en construire un intervalle de confiance pour le paramètre de pénalisation, mais il serait également intéressant de regarder si le paramètre optimal "moyen" ou "médian" augmente les capacités prédictives. En utilisant le lien entre paramètre de pénalisation et héritabilité génomique nous pourrons également regarder si nous pouvons gagner en précision pour l'estimation d'héritabilité.

Des questions plus pratiques sont également encore ouvertes : comment évolue cette distribution selon N_c ? Combien de répétitions sont nécessaires (et donc combien de matrices de contraste devons nous générer) ?

Dandine-Roulland et al. [2016] ont étudié l'utilité des composantes principales pour corriger les biais d'estimation d'héritabilité dus aux structures de population, il sera intéressant d'étudier plus en détail leurs résultats car leur idée d'intégrer les composantes principales une par une ressemble à ce que l'on voudrait faire puisque il y a un lien direct entre composantes principales et les colonnes de $\mathbf{U}$ qui sont des vecteurs propres.

7.3 Utilisation de la régression ridge hétéroscédastique

Nous avons montré et utilisé à de nombreuses reprises les liens entre régression ridge et modèles aléatoires. Dans ce manuscrit nous sommes restés concentrés sur la régression ridge avec un unique paramètre de pénalisation, ce qui correspond à un modèle à effets aléatoires avec matrice de covariance du vecteur d'effet des variants à diagonale constante. Parfois nous souhaiterions pouvoir séparer nos variants selon plusieurs

tailles d'effets (par exemple si nous souhaitons avoir un effet par chromosome), ce qui du point de vue des modèles mixtes revient à supposer que la matrice de covariance des effets n'est pas à diagonale constante. La traduction en régression ridge serait d'avoir plusieurs pénalités pour plusieurs groupes de variants : on parle alors de régression ridge hétéroscédastique. La régression ridge hétéroscédastique a permis à Hofheinz and Frisch [2014] d'obtenir de meilleures performances prédictives qu'avec une régression ridge homoscédastique sur des données de plantes. Il serait donc intéressant d'essayer sur des données humaines (de Vlaming and Groenen [2015] citait cette idée sans s'y intéresser davantage et Speed and Balding [2014] ont proposé une telle approche sur des traits binaires).

Le choix des paramètres de pénalisation dans la régression ridge hétéroscédastique pose des difficultés et des auteurs se sont déjà penchés sur la question [Hofheinz and Frisch, 2014, van de Wiel et al., 2020]. Un de mes objectifs futurs est d'adapter la GCV à la régression hétéroscédastique, et l'article de van de Wiel et al. [2020] me semble être un bon point de départ.

Un autre interêt de la régression ridge hétéroscédastique est de permettre une estimation jointe des effets génétiques et des covariables que l'on ne souhaite pas pénaliser, en imposant une pénalité nulle pour les covariables. Cela donnerait une estimation des effets fixes plus satisfaisante que notre notre solution en deux étapes. Malheureusement un terme de pénalité nulle pour des variables entraîne la singularité du problème, la matrice de pénalisation n'étant plus inversible. van de Wiel et al. [2020] propose une solution avec des projecteurs mais une solution pourrait être de passer par des pseudo-inverses. Il faudrait alors étudier la stabilité et les propriétés d'un tel estimateur.

7.4 Vérification de l'approximation de pouvoir prédictif pour des données réelles.

Nous avons montré la validité de notre méthode pour approcher différents pouvoirs prédictifs sur des simulations, mais le passage aux données réelles ne me convient pas entièrement : en particulier nous n'avons pas regardé les zones où $n \simeq p$ et $n > p$. Il serait très intéressant de regarder ces zones, mais 1) il nous faudrait une immense quantité de données et 2) nos méthodes ne peuvent tout simplement pas fonctionner pour des tailles d'échantillon aussi grandes : la complexité de l'ACP est $\mathcal{O}(n^3)$ et celle du calcul de la matrice de ressemblance $\mathcal{O}(pn^2)$, ce qui rend les temps de calcul trop

longs pour $n \simeq p \simeq 500\ 000$. Je vois donc deux approches :

— Une première idée très simple à mettre en oeuvre serait de filtrer les variants (en utilisant par exemple un filtre sur le DL des données, mais on pourrait également imaginer une sélection de variables préalable avec un dérivé du LASSO [Tibshirani, 1996], [Gorfine et al., 2017]). Le principal problème est la perte d'information, mais je pense que pour des phénotypes avec une forte héritabilité tels que la taille, les résultats seront intéressants.

— Une seconde idée serait d'approximer les matrices de ressemblance avec par exemple les approches de Nystöm [Nyström, 1930] ou d'échantillonage de colonnes. L'article de Homrighausen and McDonald [2016] me semble un bon point de départ.

7.5 Quitter le modèle linéaire avec la kernel ridge

Une extension intéressante pour ce travail serait de quitter le modèle linéaire en utilisant la *kernel ridge regression* (KRR) [Murphy, 2012], une méthode de régression dans un espace défini par des combinaisons non-nécessairement linéaires de nos variables d'entrée. Cela nous permettra en particulier de tendre vers un modèle biologique sans doute plus réaliste que le modèle linéaire, et nous pouvons donc espérer une augmentation des capacités prédictives sur les données réelles [Endelman, 2011, Morota and Gianola, 2014]. Un objectif ambitieux serait également de réussir à définir une héritabilité "non-linéaire". Nous avons déjà commencé ce travail puisque la régression ridge est en fait un cas particulier de KRR avec un noyau linéaire.

Ce passage dans un espace non-linéaire amène de nouveaux défis : quels noyaux sont les plus pertinents ? comment choisir les hyperparamètres du modèle tels que la pénalisation de la régression ridge mais également ceux associés au noyau lui-même ?

7.6 Une régression ridge à pénalisation négative

Dans la section 6.1.2 nous avions énoncé les biais des approches basées sur l'AI-REML dans le cas de phéntoypes binaires et également vu que la correction pour passer de l'héritabilité à l'échelle de l'observation à l'héritabilité à l'échelle de la liability était la même entre GCTA et PCGC. Vu que les deux corrections sont identiques, la pente de la régression dans PCGC peut être interprétée comme une héritabilité à l'échelle de l'observation. Or contrairement à GCTA où l'héritabilité à l'échelle de

l'observation est bornée entre 0 et 1, la pente de la régression peut dépasser 1 et donc l'héritabilité à l'échelle de l'observation peut être supérieure à la correction (notons qu'il est théoriquement possible pour GCTA de ne PAS contraindre l'héritabilité à l'échelle de l'observation mais je n'ai pas réussi à faire tourner l'algorithme qui tombait sur des valeurs de composantes de variance rendant la matrice Σ singulière) . Du point de vue de la ridge, une héritabilité supérieure à 1 se traduit par un paramètre de pénalisation négatif.

Je pense qu'il serait très intéressant d'utiliser la régression ridge en autorisant la pénalisation à être négative pour estimer l'héritabilité à l'échelle de l'observation, puis de réutiliser la correction de Lee et al. pour obtenir l'héritabilité à l'échelle de la liability Des questions se posent sur le comportement et la stabilité de l'erreur quand la pénalisation est négative, mais je suis confiant sur le fait que cette approche donnerait des estimations d'héritabilité moins biaisées.

Bibliographie

H. Akaike and BN Petrov ; F Csaki, editors. 2nd International Symposium on Information Theory. Akadémiai Kiadó, Budapest, Hungary, 1973. 26

C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences, 99(10) :6562–6566, May 2002. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.102102699. URL http://www.pnas.org/cgi/doi/10.1073/pnas.102102699. 59

Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. Nature Protocols, 5(9) :1564–1573, September 2010. ISSN 1754-2189, 1750-2799. doi : 10.1038/nprot.2010.116. URL http://www.nature.com/articles/nprot.2010.116. 76

S. Brard and A. Ricard. Is the use of formulae a reliable way to predict the accuracy of genomic selection ? Journal of Animal Breeding and Genetics, 132(3) :207–217, June 2015. ISSN 09312668. doi : 10.1111/jbg.12123. URL http://doi.wiley.com/10.1111/jbg.12123. 84, 85

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics, 47(3) :291–295, March 2015. ISSN 1061-4036, 1546-1718. doi : 10.1038/ng.3211. URL http://www.nature.com/articles/ng.3211. 4, 5

William S. Bush and Jason H. Moore. Chapter 11 : Genome-Wide Association Studies. PLOS Computational Biology, 8(12) :e1002822, December 2012. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1002822. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822. Publisher : Public Library of Science. 2

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on ~500,000 UK Biobank participants. preprint, Genetics, July 2017. URL http://biorxiv.org/lookup/doi/10.1101/166298. 76

Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. Frontiers in Genetics, 5, 2014. ISSN 1664-8021. doi : 10.3389/fgene.2014.00107. URL https://www.frontiersin.org/articles/10.3389/fgene.2014.00107/full. Publisher : Frontiers. 4

Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. Current Opinion in Biotechnology, 58 :161–167, August 2019. ISSN 0958-1669. doi : 10.1016/j.copbio.2019.03.004. URL http://www.sciencedirect.com/science/article/pii/S0958166918301903. 1

Samuel A Clark, John M Hickey, Hans D Daetwyler, and Julius HJ van der Werf. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution, 44(1) :4, December 2012. ISSN 1297-9686. doi : 10.1186/1297-9686-44-4. URL https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-44-4. 5

Wikipedia contributors. Truncated normal distribution, January 2020. URL https://en.wikipedia.org/w/index.php?title=Truncated_normal_distribution&oldid=935718105. Page Version ID : 935718105. 125

Hans D. Daetwyler, Beatriz Villanueva, and John A. Woolliams. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE, 3(10) : e3395, October 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0003395. URL https://dx.plos.org/10.1371/journal.pone.0003395. 84, 85, 102

Hans D. Daetwyler, Ricardo Pong-Wong, Beatriz Villanueva, and John A. Woolliams. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics, 185(3) :1021–1031, July 2010. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.110.116855. URL http://www.genetics.org/lookup/doi/10.1534/genetics.110.116855. 85, 94

Claire Dandine-Roulland. Modélisation de la composante génétique des maladies humaines : Données familiales et Modèles Mixtes. PhD thesis, Université Paris-Saclay, october 2014. 29, 31, 76, 140

Claire Dandine-Roulland and Hervé Perdry. The Use of the Linear Mixed Model in Human Genetics. Human Heredity, 80(4) :196–206, 2015. ISSN 0001-5652, 1423-0062. doi : 10.1159/000447634. URL https://www.karger.com/Article/FullText/447634. Publisher : Karger Publishers. 86

Claire Dandine-Roulland, Céline Bellenguez, Stéphanie Debette, Philippe Amouyel, Emmanuelle Génin, and Hervé Perdry. Accuracy of heritability estimations in presence of hidden population stratification. Scientific Reports, 6(1) :26471, May 2016. ISSN 2045-2322. doi : 10.1038/srep26471. URL http://www.nature.com/articles/srep26471. 5, 147

Gustavo De los Campos, Ana I. Vazquez, Rohan Fernando, Yann C. Klimentidis, and Daniel Sorensen. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genetics, 9(7) :e1003608, July 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003608. URL https://dx.plos.org/10.1371/journal.pgen.1003608. 5, 85

Ronald de Vlaming and Patrick J. F. Groenen. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Research International, 2015 :1–18, 2015. ISSN 2314-6133, 2314-6141. doi : 10.1155/2015/143712. URL http://www.hindawi.com/journals/bmri/2015/143712/. 56, 61, 62, 85, 148

Everett R. Dempster and Michael Lerner. HERITABILITY OF THRESHOLD CHARACTERS. Genetics, 35(2) :212–236, March 1950. ISSN 0016-6731. URL https://www.genetics.org/content/genetics/35/2/212.full.pdf. 124

Diabetes Genetics Initiative, The Wellcome Trust Case Control Consortium, Cambridge GEM Consortium, Michael N Weedon, Hana Lango, Cecilia M Lindgren, Chris Wallace, David M Evans, Massimo Mangino, Rachel M Freathy, John R B Perry, Suzanne Stevens, Alistair S Hall, Nilesh J Samani, Beverly Shields, Inga Prokopenko, Martin Farrall, Anna Dominiczak, Toby Johnson, Sven Bergmann, Jacques S Beckmann, Peter Vollenweider, Dawn M Waterworth, Vincent Mooser, Colin N A Palmer, Andrew D Morris, Willem H Ouwehand, Mark Caulfield, Patricia B Munroe, Andrew T Hattersley, Mark I McCarthy, and Timothy M Frayling. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics,

40(5) :575–583, May 2008. ISSN 1061-4036, 1546-1718. doi : 10.1038/ng.121. URL http://www.nature.com/articles/ng.121. 3

Jean-Michel Elsen. An analytical framework to derive the expected precision of genomic selection. Genetics Selection Evolution, 49(1) :95, December 2017. ISSN 1297-9686. doi : 10.1186/s12711-017-0366-6. URL https://gsejournal.biomedcentral.com/articles/10.1186/s12711-017-0366-6. 85

Jeffrey B. Endelman. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal, 4(3) :250, 2011. ISSN 1940-3372. doi : 10.3835/plantgenome2011.08.0024. URL https://www.crops.org/publications/tpg/abstracts/4/3/250. 149

D. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. ann hum genet 29 : 51-76. Annals of Human Genetics, 29 : 51 – 76, 09 2007. doi : 10.1111/j.1469-1809.1965.tb00500.x. 122

Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2) :399–433, 1919. 2, 54

Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1b in Europe and East Asia. Am J Hum Genet, 98(3) :456–72, March 2016. ISSN 1537-6605. doi : 10.1016/j.ajhg.2015.12.022. 76

Sophie Garnier, Magdalena Harakalova, Stefan Weiss, Michal Mokry, Vera Regitz-Zagrosek, Christian Hengstenberg, Thomas P. Cappola, Richard Isnard, Eloisa Arbustini, Stuart A. Cook, Jessica van Setten, Jörg Callis, Hakon Hakonarson, Michael P. Morley, Klaus Stark, Sanjay K. Prasad, Jin Li, Declan P. O'Regan, Maurizia Grasso, Martina Müller-Nurasyid, Thomas Meitinger, Jean-Philippe Empana, Konstantin Strauch, Mélanie Waldenberger, Kenneth B. Marguiles, Christine E. Seidman, Benjamin Meder, Pierre Boutouyrie, Patrick Lacolley, Xavier Jouven, Jeanette Erdman, Stefan Blankenberg, Thomas Wichter, Volker Ruppert, Luigi Tavazzi, Olivier Dubourg, Gerard Roizes, Richard Dorent, Pascal DeGroote, Laurent Fauchier, Jean-Noël Trochu, Jean-François Aupetit, Marine Germain, Uwe Völker, Hemerich Daiane, Ibticem Raji, Delphine Bacq-Daian, Carole Proust, Kristin Lehnert, Renee Maas, Robert Olaso, Ganapathivarma Saripella, Stephan B. Felix, Steven Mc Ginn, Laëtitia

Duboscq-Bidot, Alain van Mil, Céline Besse, Vincent Fontaine, Hélène Blanché, Brendan Keating, Pablo Garcia-Pavia, Angélique Curjol, Anne Boland, Michel Komajda, François Cambien, Jean-François Deleuze, Marcus Dörr, Folkert W. Asselbergs, Eric Villard, David-Alexandre Trégouët, Philippe Charron, and On behalf of GENMED Consortium. Genome wide association analysis in dilated cardiomyopathy reveals two new key players in systolic heart failure on chromosome 3p25.1 and 22q11.23. bioRxiv, page 2020.02.28.969147, February 2020. doi : 10.1101/2020.02.28.969147. URL https://www.biorxiv.org/content/10.1101/2020.02.28.969147v1. Publisher : Cold Spring Harbor Laboratory Section : New Results. 135

Tian Ge, Chia-Yen Chen, Benjamin M. Neale, Mert R. Sabuncu, and Jordan W. Smoller. Phenome-wide heritability analysis of the UK Biobank. PLOS Genetics, 13(4) :e1006711, April 2017. ISSN 1553-7404. doi : 10.1371/journal.pgen.1006711. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006711. Publisher : Public Library of Science. 76, 79

Mike Goddard. Genomic selection : prediction of accuracy and maximisation of long term response. Genetica, 136(2) :245–257, June 2009. ISSN 0016-6707, 1573-6857. doi : 10.1007/s10709-008-9308-0. URL http://link.springer.com/10.1007/s10709-008-9308-0. 1, 85, 94

David Golan, Eric S. Lander, and Saharon Rosset. Measuring missing heritability : Inferring the contribution of common variants. Proceedings of the National Academy of Sciences, 111(49) :E5272–E5281, December 2014. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1419064111. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1419064111. 4, 61, 125, 126, 127, 128, 130, 131, 134, 144, XXV

G. Golub and W. Kahan. Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis, 2(2) :205–224, January 1965. ISSN 0887-459X. doi : 10.1137/0702016. URL http://epubs.siam.org/doi/10.1137/0702016. 19

G. H. Golub and C. Reinsch. Singular Value Decomposition and Least Squares Solutions. Numer. Math., 14(5) :403–420, April 1970. ISSN 0029-599X. doi : 10.1007/BF02163027. URL https://doi.org/10.1007/BF02163027. Place : Berlin, Heidelberg Publisher : Springer-Verlag. 19

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as

a Method for Choosing a Good Ridge Parameter. Technometrics, 21(2) :215–233, 1978. 28

Malka Gorfine, Sonja I. Berndt, Jenny Chang-Claude, Michael Hoffmeister, Loic Le Marchand, John Potter, Martha L. Slattery, Nir Keret, Ulrike Peters, and Li Hsu. Heritability Estimation using a Regularized Regression Approach (HERRA) : Applicable to continuous, dichotomous or age-at-onset outcome. PLOS ONE, 12(8) : e0181269, August 2017. ISSN 1932-6203. doi : 10.1371/journal.pone.0181269. URL https://dx.plos.org/10.1371/journal.pone.0181269. 149

Emmanuelle Génin. Missing heritability of complex diseases : case solved ? Human Genetics, June 2019. ISSN 0340-6717, 1432-1203. doi : 10.1007/s00439-019-02034-4. URL http://link.springer.com/10.1007/s00439-019-02034-4. 3

Margaret A. Hamburg and Francis S. Collins. The Path to Personalized Medicine. New England Journal of Medicine, 363(4) :301–304, July 2010. ISSN 0028-4793. doi : 10.1056/NEJMp1006304. URL https://doi.org/10.1056/NEJMp1006304. Publisher : Massachusetts Medical Society _eprint : https ://doi.org/10.1056/NEJMp1006304. 1

J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics, 2(1) :3–19, March 1972. ISSN 1573-3297. doi : 10.1007/BF01066731. URL https://doi.org/10.1007/BF01066731. 4, 126

B.J. Hayes, P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. Invited review : Genomic selection in dairy cattle : Progress and challenges. Journal of Dairy Science, 92(2) :433–443, February 2009. ISSN 00220302. doi : 10.3168/jds.2008-1646. URL https://linkinghub.elsevier.com/retrieve/pii/S0022030209703479. 1, 5

Elliot L. Heffner, Mark E. Sorrells, and Jean-Luc Jannink. Genomic Selection for Crop Improvement. Crop Science, 49(1) :1–12, 2009. ISSN 1435-0653. doi : 10.2135/cropsci2008.08.0512. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.2135/cropsci2008.08.0512. _eprint : https ://acsess.onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2008.08.0512. 1

Joel N. Hirschhorn and Mark J. Daly. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6(2) :95–108, February 2005. ISSN 1471-0064. doi : 10.1038/nrg1521. URL https://www.nature.com/articles/nrg1521. Number : 2 Publisher : Nature Publishing Group. 2

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression : Biased Estimation for Nonorthogonal Problems. Technometrics, 42(1) :80, February 2000. ISSN 00401706. doi : 10.2307/1271436. URL https://www.jstor.org/stable/1271436?origin=crossref. 17

Nina Hofheinz and Matthias Frisch. Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation. G3: Genes|Genomes|Genetics, 4(3) :539–546, March 2014. ISSN 2160-1836. doi : 10.1534/g3.113.010025. URL http://g3journal.org/lookup/doi/10.1534/g3.113.010025. 148

Darren Homrighausen and Daniel J. McDonald. On the Nystr\"om and Column-Sampling Methods for the Approximate Principal Components Analysis of Large Data Sets. Journal of Computational and Graphical Statistics, 25(2) :344–362, April 2016. ISSN 1061-8600, 1537-2715. doi : 10.1080/10618600.2014.995799. URL http://arxiv.org/abs/1602.01120. arXiv : 1602.01120. 149

H. Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6) :417–441, 1933. ISSN 1939-2176(Electronic),0022-0663(Print). doi : 10.1037/h0071325. Place : US Publisher : Warwick & York. 20

Kangcheng Hou, Kathryn S. Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. Nature Genetics, 51(8) :1244–1251, August 2019. ISSN 1061-4036, 1546-1718. doi : 10.1038/s41588-019-0465-0. URL http://www.nature.com/articles/s41588-019-0465-0. 4

M. Höge. On the Way to Appropriate Model Complexity. AGU Fall Meeting Abstracts, 13, December 2016. URL http://adsabs.harvard.edu/abs/2016AGUFMNG13A1683H. 13

Kewal K. Jain. Personalized medicine. Current Opinion in Molecular Therapeutics, 4 (6) :548–558, December 2002. ISSN 1464-8431. 1

Emre Karaman, Hao Cheng, Mehmet Z. Firat, Dorian J. Garrick, and Rohan L. Fernando. An Upper Bound for Accuracy of Prediction Using GBLUP. PLOS ONE,

11(8) :e0161054, August 2016. ISSN 1932-6203. doi : 10.1371/journal.pone.0161054. URL http://dx.plos.org/10.1371/journal.pone.0161054. 5

Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Estimating Missing Heritability for Disease from Genome-wide Association Studies. The American Journal of Human Genetics, 88(3) :294–305, March 2011. ISSN 00029297. doi : 10.1016/j.ajhg.2011.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929711000206. 4, 122, 123, 128, 137, 138, 150

C. L. Mallows. Some comments on c p. Technometrics, 15(4) :661–675, 1973. doi : 10.1080/00401706.1973.10489103. URL https://doi.org/10.1080/00401706.1973.10489103. 26

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. Nature, 461(7265) : 747–753, October 2009. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature08494. URL http://www.nature.com/doifinder/10.1038/nature08494. 3

Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits : consensus, uncertainty and challenges. Nature Reviews Genetics, 9(5) :356–369, May 2008. ISSN 1471-0064. doi : 10.1038/nrg2344. URL https://www.nature.com/articles/nrg2344. Number : 5 Publisher : Nature Publishing Group. 2

Rosa J. Meijer and Jelle J. Goeman. Efficient approximate k -fold and leave-one-out cross-validation for ridge regression : Efficient approximate k -fold and leave-one-out cross-validation. Biometrical Journal, 55(2) :141–155, March 2013. ISSN 03233847. doi : 10.1002/bimj.201200088. URL http://doi.wiley.com/10.1002/bimj.201200088. 27

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics, 157(4) :1819–1829, April

2001. ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/157/4/1819. Publisher : Genetics Section : Investigations. 5

Gota Morota and Daniel Gianola. Kernel-based whole-genome prediction of complex traits : a review. Frontiers in Genetics, 5, 2014. ISSN 1664-8021. doi : 10.3389/fgene.2014.00363. URL https://www.frontiersin.org/articles/10.3389/fgene.2014.00363/full. Publisher : Frontiers. 149

Kevin P Murphy. Machine Learning : a probalistic perspective. The MIT Press, the mit press edition, 2012. URL https://mitpress.mit.edu/books/machine-learning-1. Library Catalog : mitpress.mit.edu. 149

E. J. Nyström. Über Die Praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. Acta Mathematica, 54 :185–204, 1930. ISSN 0001-5962, 1871-2509. doi : 10.1007/BF02547521. URL https://projecteuclid.org/euclid.acta/1485887849. Publisher : Institut Mittag-Leffler. 149

Karl Pearson. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11) :559–572, November 1901. ISSN 1941-5982, 1941-5990. doi : 10.1080/14786440109462720. URL https://www.tandfonline.com/doi/full/10.1080/14786440109462720. 20

Hervé Perdry. Héritabilité : de la régression vers la moyenne au modèle mixte. Mémoire présenté en vue de l'obtention du diplôme d'Habilitation à Diriger les Recherches, Université Paris-Sud, 2017. 29

Hervé Perdry and Claire Dandine-Roulland. gaston : Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models, 2017. URL https://CRAN.R-project.org/package=gaston. R package version 1.5. 72

Paul D.P. Pharoah, Antonis Antoniou, Martin Bobrow, Ron L. Zimmern, Douglas F. Easton, and Bruce A.J. Ponder. Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics, 31(1) :33–36, May 2002. ISSN 1061-4036, 1546-1718. doi : 10.1038/ng853. URL http://www.nature.com/articles/ng853z. 84

Alkes L. Price, Michael E. Weale, Nick Patterson, Simon R. Myers, Anna C. Need, Kevin V. Shianna, Dongliang Ge, Jerome I. Rotter, Esther Torres, Kent D. Taylor, David B. Goldstein, and David Reich. Long-Range LD Can Confound Genome

Scans in Admixed Populations. The American Journal of Human Genetics, 83(1) : 132–135, July 2008. ISSN 00029297. doi : 10.1016/j.ajhg.2008.06.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929708003534. 76

Shaun Purcell. PLINK 1.9, 2009. URL https://www.cog-genomics.org/plink2/. 135

Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, Pamela Sklar, Shaun M. Purcell (Leader), Jennifer L. Stone, Patrick F. Sullivan, Douglas M. Ruderfer, Andrew McQuillin, Derek W. Morris, Colm T. O'Dushlaine, Aiden Corvin, Peter A. Holmans, Michael C. O'Donovan, Pamela Sklar, Naomi R. Wray, Stuart Macgregor, Pamela Sklar, Patrick F. Sullivan, Michael C. O'Donovan, Peter M. Visscher, Hugh Gurling, Douglas H. R. Blackwood, Aiden Corvin, Nick J. Craddock, Michael Gill, Christina M. Hultman, George K. Kirov, Paul Lichtenstein, Andrew McQuillin, Walter J. Muir, Michael C. O'Donovan, Michael J. Owen, Carlos N. Pato, Shaun M. Purcell, Edward M. Scolnick, David St Clair, Jennifer L. Stone, Patrick F. Sullivan, Pamela Sklar (Leader), Michael C. O'Donovan, George K. Kirov, Nick J. Craddock, Peter A. Holmans, Nigel M. Williams, Lyudmila Georgieva, Ivan Nikolov, N. Norton, H. Williams, Draga Toncheva, Vihra Milanova, Michael J. Owen, Christina M. Hultman, Paul Lichtenstein, Emma F. Thelander, Patrick Sullivan, Derek W. Morris, Colm T. O'Dushlaine, Elaine Kenny, Emma M. Quinn, Michael Gill, Aiden Corvin, Andrew McQuillin, Khalid Choudhury, Susmita Datta, Jonathan Pimm, Srinivasa Thirumalai, Vinay Puri, Robert Krasucki, Jacob Lawrence, Digby Quested, Nicholas Bass, Hugh Gurling, Caroline Crombie, Gillian Fraser, Soh Leh Kuan, Nicholas Walker, David St Clair, Douglas H. R. Blackwood, Walter J. Muir, Kevin A. McGhee, Ben Pickard, Pat Malloy, Alan W. Maclean, Margaret Van Beck, Naomi R. Wray, Stuart Macgregor, Peter M. Visscher, Michele T. Pato, Helena Medeiros, Frank Middleton, Celia Carvalho, Christopher Morley, Ayman Fanous, David Conti, James A. Knowles, Carlos Paz Ferreira, Antonio Macedo, M. Helena Azevedo, Carlos N. Pato, Jennifer L. Stone, Douglas M. Ruderfer, Andrew N. Kirby, Manuel A. R. Ferreira, Mark J. Daly, Shaun M. Purcell, Pamela Sklar, Shaun M. Purcell, Jennifer L. Stone, Kimberly Chambert, Douglas M. Ruderfer, Finny Kuruvilla, Stacey B. Gabriel, Kristin Ardlie, Jennifer L. Moran, Mark J. Daly, Edward M. Scolnick, Pamela Sklar, The International Schizophrenia Consortium, Manuscript preparation, Data analysis, GWAS analysis subgroup, Polygene analyses subgroup, Management commit-

tee, Cardiff University, Karolinska Institutet/University of North Carolina at Chapel Hill, Trinity College Dublin, University College London, University of Aberdeen, University of Edinburgh, Queensland Institute of Medical Research, University of Southern California, Massachusetts General Hospital, and Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460 (7256) :748–752, August 2009. ISSN 1476-4687. doi : 10.1038/nature08185. URL https://www.nature.com/articles/nature08185. Number : 7256 Publisher : Nature Publishing Group. 84

Charles-Elie Rabier, Philippe Barre, Torben Asp, Gilles Charmet, and Brigitte Mangin. On the Accuracy of Genomic Selection. PLOS ONE, 11(6) :e0156086, June 2016. ISSN 1932-6203. doi : 10.1371/journal.pone.0156086. URL https://dx.plos.org/10.1371/journal.pone.0156086. 5, 85, 102

Doug Speed and David J. Balding. MultiBLUP : improved SNP-based prediction for complex traits. Genome Research, 24(9) :1550–1557, January 2014. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.169375.113. URL http://genome.cshlp.org/content/24/9/1550. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab. 5, 148

Doug Speed, Gibran Hemani, Michael R. Johnson, and David J. Balding. Improved Heritability Estimation from Genome-wide SNPs. The American Journal of Human Genetics, 91(6) :1011–1021, December 2012. ISSN 00029297. doi : 10.1016/j.ajhg.2012.10.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929712005332. 4

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine, 12(3) : e1001779, March 2015. ISSN 1549-1676. doi : 10.1371/journal.pmed.1001779. URL https://dx.plos.org/10.1371/journal.pmed.1001779. 73

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the

Royal Statistical Society. Series B (Methodological), 58(1) :267–288, 1996. URL http://www.jstor.org/stable/2346178. 16, 149

A N Tikhonov. On the solution of ill-posed problems and the method of regularization. SIAM Rev, page 5, 1965. 17

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer, 2nd edition edition, February 2009. 14, 27

Mark A. van de Wiel, Mirrelijn M. van Nee, and Armin Rauschenberger. Fast cross-validation for multi-penalty ridge regression. arXiv :2005.09301 [stat], May 2020. URL http://arxiv.org/abs/2005.09301. arXiv : 2005.09301. 148

Peter M. Visscher. Sizing up human height variation. Nature Genetics, 40(5) :489–490, May 2008. ISSN 1546-1718. doi : 10.1038/ng0508-489. URL https://www.nature.com/articles/ng0508-489. Number : 5 Publisher : Nature Publishing Group. 3

Peter M. Visscher and Michael E. Goddard. A General Unified Framework to Assess the Sampling Variance of Heritability Estimates Using Pedigree or Marker-Based Relationships. Genetics, 199(1) :223–232, January 2015. ISSN 0016-6731, 1943-2631. doi : 10.1534/genetics.114.171017. URL http://www.genetics.org/lookup/doi/10.1534/genetics.114.171017. 66, 140

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7) : 565–569, July 2010. ISSN 1061-4036, 1546-1718. doi : 10.1038/ng.608. URL http://www.nature.com/articles/ng.608. 3, 52, 53, 55, 74, 101, 126, 137, 141

Bingxin Zhao and Hongtu Zhu. Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. arXiv :1911.10142 [stat], November 2019. URL http://arxiv.org/abs/1911.10142. arXiv : 1911.10142. 86, 97

Annexe A

Preuves pour la régression ridge

A.1 Espérance d'une forme quadratique

Soit $\mathbf{y}$ une variable aléatoire avec $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ et $\text{var}(\mathbf{y}) = \boldsymbol{\Sigma}$ et $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{y}^T\mathbf{A}\mathbf{y}\right] &= \mathbb{E}\left[(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{y}-\boldsymbol{\mu}) + 2\boldsymbol{\mu}^T\mathbf{A}\mathbf{y} - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\right] \\
&= \mathbb{E}\left[(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{y}-\boldsymbol{\mu})\right] + 2\boldsymbol{\mu}^T\mathbf{A}\mathbb{E}\left[\mathbf{y}\right] - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \\
&= \mathbb{E}\left[\text{tr}\left((\mathbf{y}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{y}-\boldsymbol{\mu})\right)\right] + 2\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \\
&= \text{tr}\left(\mathbf{A}\mathbb{E}\left[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\right]\right) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \\
&= \text{tr}\left(\mathbf{A}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}
\end{aligned}$$

A.2 Décomposition biais-variance de l'erreur de prédiction

Ici les variables aléatoires sont la réponse de test ainsi que les données et la réponse de train. L'espérance est donc par rapport à ces variables.

$$\mathbb{E}_{X=x_0}\left[(Y-\hat{f}(x_0))^2\right] = \mathbb{E}_{X=x_0}\left[Y^2 - 2Y\hat{f}(x_0) + \hat{f}(x_0)^2\right] \quad (*)$$

En utilisant la formule de la variance on a

— $\mathbb{E}_{X=x_0}\left[Y^2\right] = \mathbb{E}_{X=x_0}\left[(Y - \mathbb{E}_{X=x_0}[Y])^2\right] + \mathbb{E}_{X=x_0}\left[Y\right]^2 = \text{var}(Y) + \mathbb{E}_{X=x_0}\left[Y\right]^2$,

— $\mathbb{E}_{X=x_0}\left[\hat{f}(x_0)^2\right] = \mathbb{E}_{X=x_0}\left[\left(\hat{f}(x_0) - \mathbb{E}_{X=x_0}[\hat{f}(x_0)]\right)^2\right] + \mathbb{E}_{X=x_0}\left[\hat{f}(x_0)\right]^2 = \text{var}(\hat{f}(x_0)) +$

$\mathbb{E}_{X=x_0}\left[\hat{f}(x_0)\right]^2$.

Alors

$$
\begin{aligned}
(*) &= \operatorname{var}(Y) + \mathbb{E}_{X=x_0}[Y]^2 + \operatorname{var}(\hat{f}(x_0)) + \mathbb{E}_{X=x_0}\left[\hat{f}(x_0)\right]^2 - 2\,\mathbb{E}_{Y,X=x_0}[Y]\,\mathbb{E}_{\text{train},X=x_0}\left[\hat{f}(x_0)\right] \\
&= \underbrace{\operatorname{var}(Y)}_{\text{erreur irréductible}} + \underbrace{\operatorname{var}(\hat{f}(x_0))}_{\text{variance estimateur}} + \underbrace{\left(\mathbb{E}_{X=x_0}\left[\hat{f}(x_0)\right] - \mathbb{E}_{X=x_0}[Y]\right)^2}_{\text{biais}}
\end{aligned}
$$

A.3 Théorème de l'existence

On rappelle la forme de l'estimateur des moindres carrées et celui de la régression ridge (on supposera également que la matrice $\mathbf{Z}^T\mathbf{Z}$ est inversible pour pouvoir définir l'estimateur des moindres carrés) :

$$
\begin{aligned}
\hat{u} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}, \\
\hat{u}_R &= (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p)^{-1}\mathbf{Z}^T\mathbf{y}.
\end{aligned}
$$

On remarque que si $\lambda = 0$, alors $\hat{u} = \hat{u}_R$. On va essayer de calculer le comportement de $\partial_\lambda \mathrm{MSE}(\hat{u}_R, u)$.

On utilise la décomposition biais-variance et la décomposition en valeur singulière $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T$,

$$
\begin{aligned}
\mathrm{MSE}(\hat{u}_R, u) &= \operatorname{var}(\hat{u}_R) + \operatorname{biais}^2(\hat{u}_R) \\
&= \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} + \lambda^2 \sum_{j=1}^{p} \frac{c_j^2}{(d_j^2+\lambda)^2} \text{ avec } c = \mathbf{V}^T u.
\end{aligned}
$$

On peut alors en déduire la dérivée du MSE

$$
\begin{aligned}
\partial_\lambda \mathrm{MSE}(\hat{u}_R, u) &= \partial_\lambda \operatorname{var}(\hat{u}_R) + \partial_\lambda \operatorname{biais}^2(\hat{u}_R) \\
&= \sum_{j=1}^{p} \frac{-2\sigma^2 d_j^2}{(d_j^2+\lambda)^3} + \sum_{j=1}^{p} \frac{2\lambda d_j^2 c_j^2}{(d_j^2+\lambda)^3} \\
&= \sum_{j=1}^{p} \frac{2d_j^2\left(\lambda c_j^2 - \sigma^2\right)}{(d_j^2+\lambda)^3}
\end{aligned}
$$

Ainsi si on note $c^2_{min} = \min_{j\in[\![1,p]\!]} c^2_j$, on a $\forall\lambda < \frac{\sigma^2}{c^2_{min}},\ \partial_\lambda \text{MSE}(\hat{u}_R, u) < 0$. On a donc montré que $\exists\lambda\, 0,\ \text{MSE}(\hat{u}_R, u) \leq \text{MSE}(\hat{u}, u)$.

A.4 Élements de preuves pour l'estimateur de la LOO

Dans cette section nous présenterons des éléments pour la démonstration de la formule 2.16.

La formule de Sherman-Morrison-Woodbury nous dit que pour $\mathbf{A} \in \mathcal{M}_p$ une matrice non-singulière et deux vecteurs $u, v \in \mathbb{R}^p$, on a

$$\left(\mathbf{A} + uv^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}uv^T\mathbf{A}^{-1}}{1 + v^T\mathbf{A}^{-1}u}. \tag{A.1}$$

En utilisant A.1, on peut alors écrire

$$\begin{aligned}\left(\mathbf{Z}^T_{-i}\mathbf{Z}_{-i} + \lambda\mathbf{I}_p\right)^{-1} &= \left(\mathbf{Z}^T\mathbf{Z} - z_i z_i^T + \lambda\mathbf{I}_p\right)^{-1} \\ &= \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} + \frac{\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i z_i^T \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}}{1 - z_i^T\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i}\end{aligned}$$

avec $z_i \in \mathbb{R}^p$ qui correspond à la i-ème ligne de $\mathbf{Z}$ donc à l'individu i. Si on remarque que l'on a

$$\mathbf{Z}^T_{-i}\mathbf{y}_{-i} = \mathbf{Z}^T\mathbf{y} - z_i^T y_i,$$

alors on peut écrire

$$\begin{aligned}\hat{u}_R^{-i} &= \left(\mathbf{Z}_{-i}^T\mathbf{Z}_{-i} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{Z}_{-i}^T\mathbf{y}_{-i} \\ &= \left(\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} + \frac{\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i z_i^T \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}}{1 - z_i^T\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i}\right)\left(\mathbf{Z}^T\mathbf{y} - z_i y_i\right) \\ &= \hat{u}_R - \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i y_i + \frac{\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i z_i^T \left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}}{1 - z_i^T\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i}\left(\mathbf{Z}^T\mathbf{y} - z_i y_i\right).\end{aligned}$$

En mettant au même dénominateur, en n'oubliant pas pas y_i est un scalaire et en remarquant que $z_i\left(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1} z_i^T = [h_\lambda]_{ii}$, on arrive a montrer que

$$\hat{u}_R^{-i} = \hat{u}_R - \frac{(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p)^{-1} z_i (y_i - z_i^T\hat{u}_R)}{1 - [h_\lambda]_{ii}}.$$

A.5 Élements de preuves pour la GCV

Pour la démonstration on introduit le concept de matrice circulante. Une matrice $\mathbf{C}$ est dite circulante si elle est de la forme

$$\mathbf{C} = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & & c_{n-3} \\ \vdots & & & \ddots & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{pmatrix} \in \text{Circ}(n)?$$

En particulier on voit que les matrices circulantes sont à diagonale constante.

Nous admettrons que les matrices circulantes doivent vérifier la propriété suivante : soit $\mathbf{W} \in \mathcal{M}_n(\mathbb{C})$ tel que $[\mathbf{W}]_{jk} = \frac{1}{\sqrt{n}} e^{2\pi i jk/n}$ avec $j, k \in \{1, ..., n\}$. Alors $\mathbf{W}$ diagonalise toute les matrices circulantes i.e

$$\forall \mathbf{C} \in \text{Circ}(n),\ \exists \boldsymbol{\mathcal{D}} \in \mathbb{D}_n(\mathbb{C})\ /\ \mathbf{C} = \mathbf{W}\boldsymbol{\mathcal{D}}\mathbf{W}^*$$

avec * l'opérateur de la transposée complexe ($\mathbf{X}^* = \bar{X}^*$).

L'idée de la GCV est de projeter notre modèle dans un espace complexe bien choisir pour avoir une matrice $\mathbf{H}_\lambda$ à coefficients constant. Si on note $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ la SVD de $\mathbf{Z}$, on peut alors multiplier par la droite 2.1 par $\mathbf{W}\mathbf{U}^T$

$$\begin{aligned}\mathbf{W}\mathbf{U}^T\mathbf{y} &= \mathbf{W}\mathbf{U}^T\mathbf{Z}u + \mathbf{W}\mathbf{U}^T\mathbf{e} \\ &= \mathbf{W}\mathbf{D}\mathbf{V}^T u + \mathbf{W}\mathbf{U}^T\mathbf{e}\end{aligned}$$

Puisque $\mathbf{U} \in \mathcal{O}_n(\mathbb{R})$ et $\mathbf{W} \in \mathcal{O}_n(\mathbb{C})$, on a

$$\begin{aligned}\left\|\mathbf{W}\mathbf{U}^T\mathbf{y} - \mathbf{W}\mathbf{U}^T\mathbf{Z}u\right\|_2^2 + \lambda \|u\|_2^2 &= (\mathbf{y} - \mathbf{Z}u)^T \mathbf{U}\mathbf{W}^*\mathbf{W}\mathbf{U}^T (\mathbf{y} - \mathbf{Z}u) + \lambda \|u\|_2^2 \\ &= \|\mathbf{y} - \mathbf{Z}u\|_2^2 + \lambda \|u\|_2^2\end{aligned}$$

donc le λ_{opt} est le même dans les deux modèles.

La matrice de lissage dans le nouveau modèle vaut

$$\begin{aligned}\tilde{\mathbf{H}}_\lambda &= (\mathbf{W}\mathbf{U}^T\mathbf{Z})(\mathbf{W}\mathbf{U}^T\mathbf{Z})^* \left((\mathbf{W}\mathbf{U}^T\mathbf{Z})(\mathbf{W}\mathbf{U}^T\mathbf{Z})^* + \lambda\mathbf{I}_n\right)^{-1} \\ &= (\mathbf{W}\mathbf{D}\mathbf{V}^T)(\mathbf{W}\mathbf{D}\mathbf{V}^T)^* \left((\mathbf{W}\mathbf{D}\mathbf{V}^T)(\mathbf{W}\mathbf{D}\mathbf{V}^T)^* + \lambda\mathbf{W}\mathbf{W}^*\right)^{-1} \\ &= \mathbf{W}\mathbf{D}\mathbf{D}^T(\mathbf{D}\mathbf{D}^T + \lambda\mathbf{I}_n)^{-1}\mathbf{W}^* \in \mathrm{Circ}(n)\end{aligned}$$

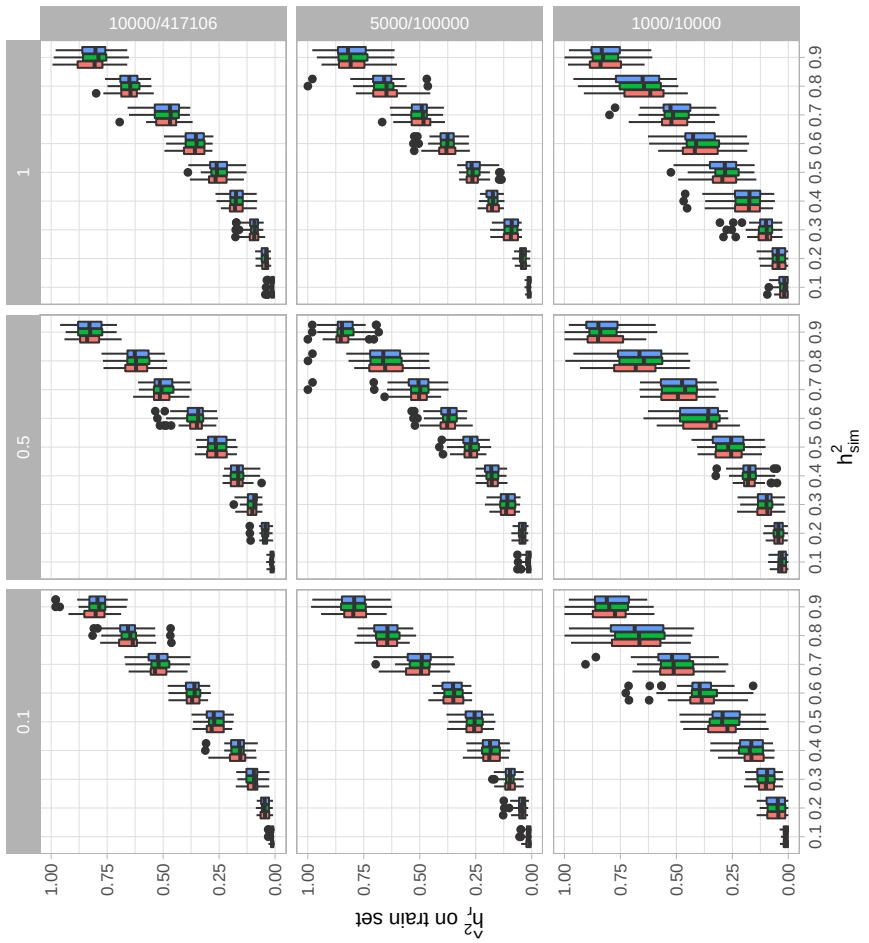
On a donc une matrice de lissage a coefficients constant ce qui donne

$$\forall i \in [\![1, n]\!],\ [h_\lambda]_{ii} = \frac{1}{n}\mathrm{tr}(\tilde{\mathbf{H}}_\lambda) = \sum_{k=1}^{n} \frac{d_k}{d_k + \lambda}.$$

En combinant ces résultats et la formule 2.17, on retrouve 2.19.

Annexe B

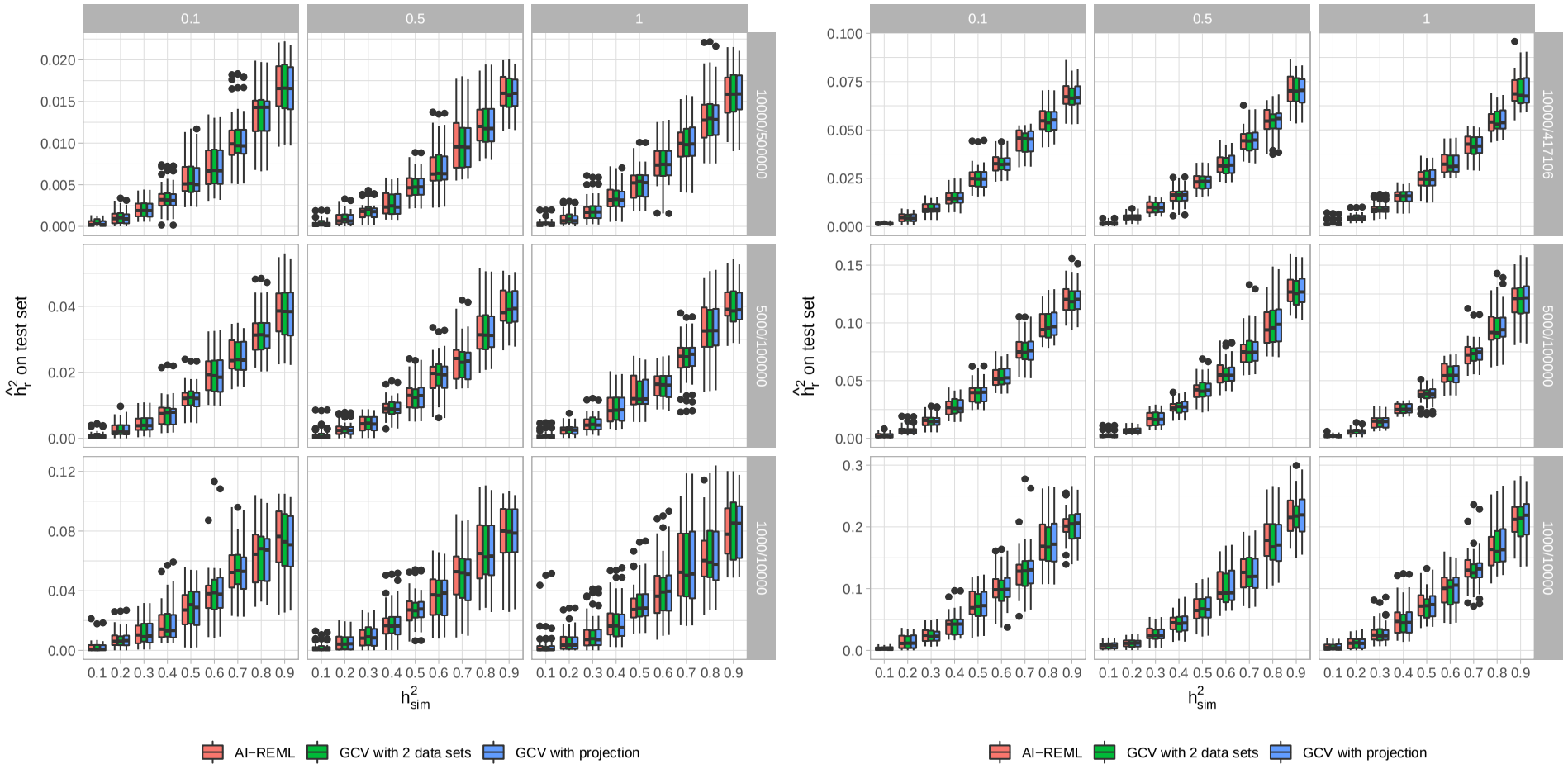
Graphes supplémentaires pour l'influence du pourcentage de variant causaux pour l'estimation d'héritabilité avec le R^2 et h_r^2



(a) Simulations synthétiques

(b) Simulations semis-synthétiques

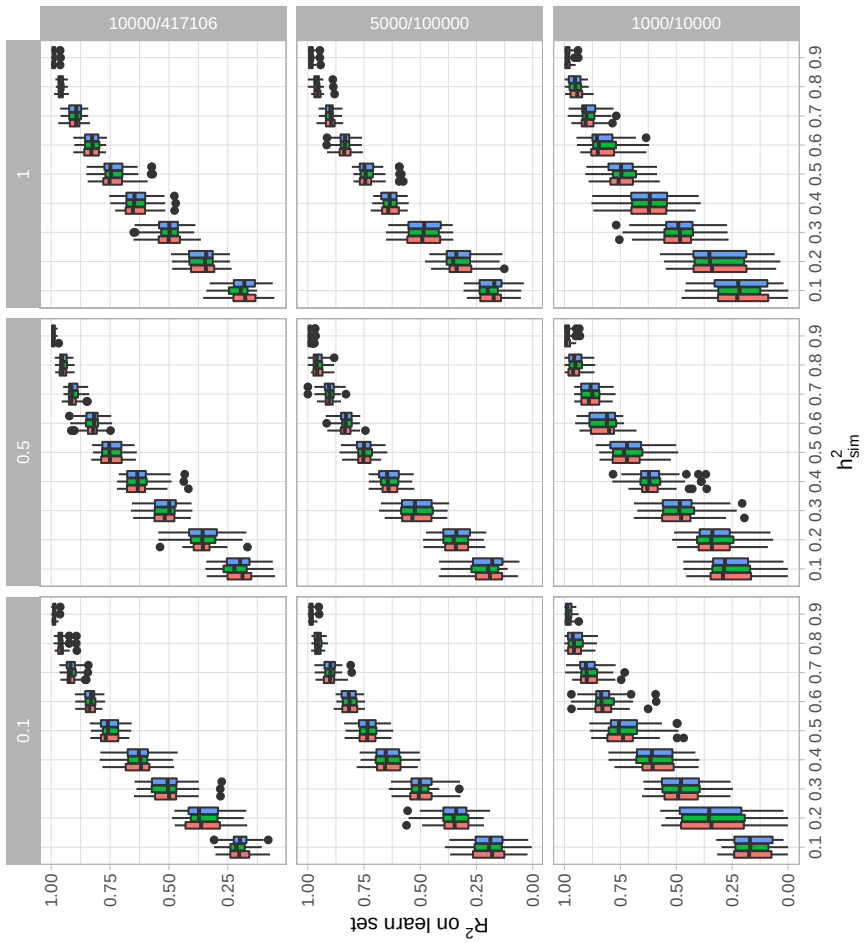
FIGURE B.1 – Graphe d'estimation de h_r^2 sur l'ensemble d'apprentissage pour les simulations. La figure de gauche correspond aux simulations synthétiques et celle de droite aux simulations semi-synthétiques. Dans chaque figure on trouve 9 graphes rangés en colonne selon f_c et en ligne selon n/p. Chacun de ces graphes est un ensemble de boxplots dont l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h_r^2 calculée sur l'ensemble d'apprentissage. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.



(a) Simulations synthétiques

(b) Simulations semis-synthétiques

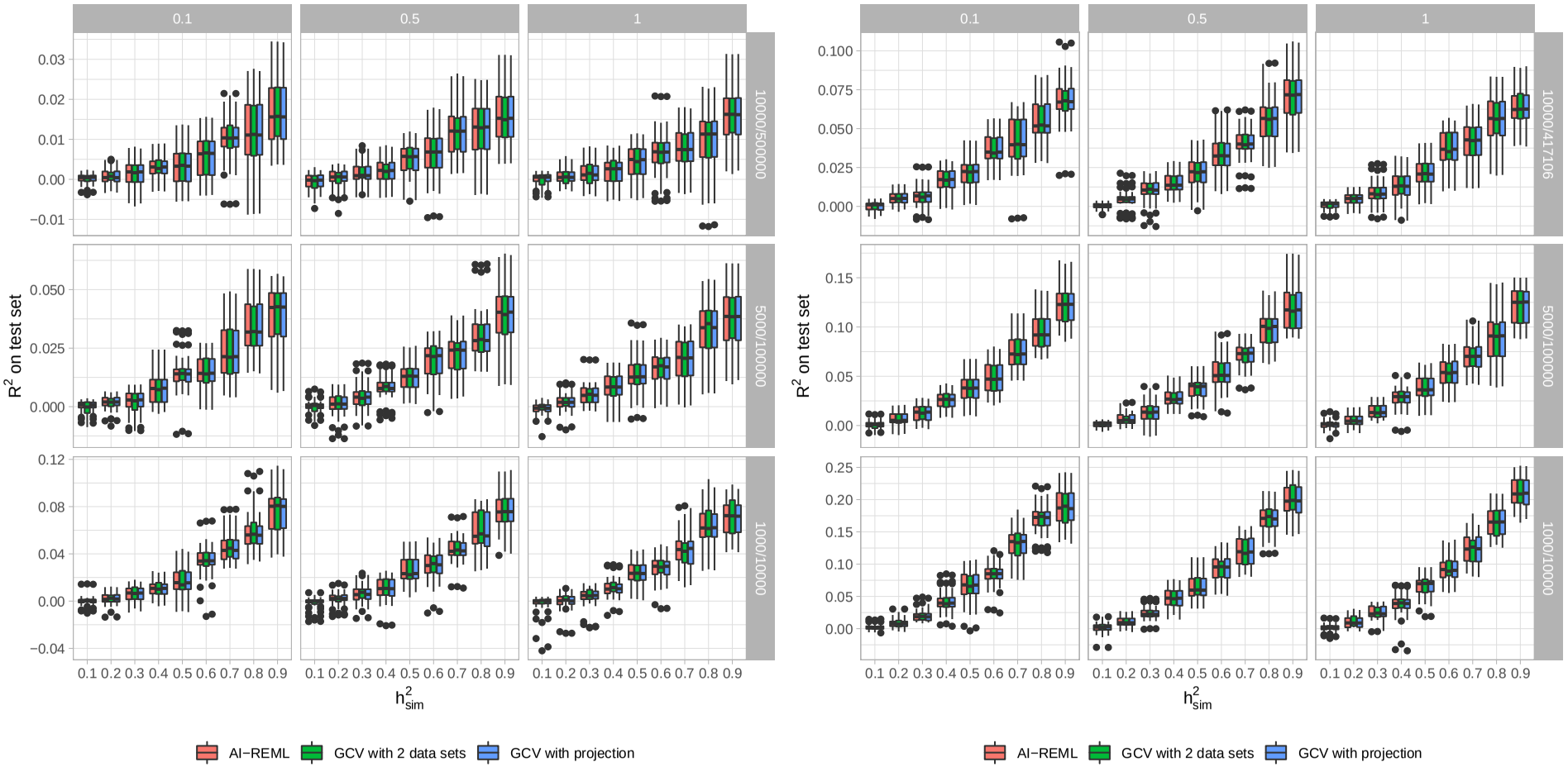
FIGURE B.2 – Graphe d'estimation de h^2_r sur l'ensemble de test pour les simulations. La figure de gauche correspond aux simulations synthétiques et celle de droite aux simulations semi-synthétiques. Dans chaque figure on trouve 9 graphes rangés en colonne selon f_c et en ligne selon n/p. Chacun de ces graphes est un ensemble de boxplots dont l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h^2_r calculée sur l'ensemble de test. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.



(a) Simulations synthétiques

(b) Simulations semis-synthétiques

FIGURE B.3 – Graphe d'estimation de h_p^2 (confondue avec le R^2) sur l'ensemble d'apprentissage pour les simulations. La figure de gauche correspond aux simulations synthétiques et celle de droite aux simulations semi-synthétiques. Dans chaque figure on trouve 9 graphes rangés en colonne selon f_c et en ligne selon n/p. Chacun de ces graphes est un ensemble de boxplots dont l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h_p^2 calculée sur l'ensemble d'apprentissage. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.



(a) Simulations synthétiques

(b) Simulations semis-synthétiques

FIGURE B.4 – Graphe d'estimation de h_p^2 (confondue avec le R^2) sur l'ensemble de test pour les simulations. La figure de gauche correspond aux simulations synthétiques et celle de droite aux simulations semi-synthétiques. Dans chaque figure on trouve 9 graphes rangés en colonne selon f_c et en ligne selon n/p. Chacun de ces graphes est un ensemble de boxplots dont l'axe des abscisses correspond à l'héritabilité simulée et l'axe des ordonnées à h_p^2 calculée sur l'ensemble de test. Les boites des boxplots correspondent respectivement aux estimations faites avec la régression ridge en utilisant la GCV corrigée avec une matrice de contraste, la régression ridge en utilisant la GCV avec un deuxième jeu de données et un BLUP dont les paramètres de variance ont été estimés avec l'AI-REML.

Annexe C

Preuves et graphes supplémentaires pour l'approximation du pouvoir prédictif

Dans cette annexe nous détaillerons les calculs pour des différents pouvoir prédictifs (MSE et corr^2) calculés sur les ensembles de test et d'apprentissage, ainsi que les calculs pour la variance empirique des coefficients du vecteur $\hat{u}_R$.

Rappelons que nous notons $\mathcal{A}$ l'ensemble d'apprentissage et $\mathcal{T}$ l'ensemble de test. Pour simplifier l'écriture, nous nous autoriserons les notations $\mathbb{E}_{\mathcal{A}} = \mathbb{E}_{\mathbf{y}_{tr}}$ et $\mathbb{E}_{\mathcal{T}} = \mathbb{E}_{y_{te},z_{te}}$.

C.1 Calcul du MSE sur un individu de test

C.1.1 Deux égalités

$$\frac{n}{n+\lambda} = \frac{n}{n+p\frac{1-h^2}{h^2}} = \frac{\frac{n}{p}}{\frac{n}{p}+\frac{1-h^2}{h^2}} = \frac{\frac{n}{p}\times h^2}{n/p\times h^2+(1-h^2)} = \frac{\frac{n}{p}\times h^2}{1+h^2\times(\frac{n}{p}-1)}$$

$$\frac{\lambda}{n+\lambda} = \frac{p\frac{1-h^2}{h^2}}{n+p\frac{1-h^2}{h^2}} = \frac{1-h^2}{\frac{n}{p}\times h^2+(1-h^2)} = \frac{1-h^2}{1+h^2(\frac{n}{p}-1)}$$

C.1.2 Limites des quantités

Le carré du biais

$$\lim_{\frac{n}{p}\to 1-} h^2\left(1+\frac{n}{p}\left((h^2)^2-2h^2\right)\right) = \lim_{\frac{n}{p}\to 1+}\left(\frac{1-h^2}{1+h^2(\frac{n}{p}-1)}\right)^2 h^2 = (1-h^2)^2h^2$$

La variance

$$\lim_{\frac{n}{p}\to 1-}(1-h^2)(h^2)^2\frac{n}{p} = \lim_{\frac{n}{p}\to 1+}\sigma^2\frac{1}{\frac{n}{p}}\left(\frac{\frac{n}{p}\times h^2}{1+h^2\times(\frac{n}{p}-1)}\right)^2 = (1-h^2)(h^2)^2$$

L'erreur quadratique

$$\lim_{\frac{n}{p}\to 1-} 1-\frac{n}{p}(h^2)^2 = \lim_{\frac{n}{p}\to 1+}(1-h^2)\frac{1+\frac{n}{p}h^2}{1+h^2(\frac{n}{p}-1)} = 1-(h^2)^2$$

C.2 Calcul du MSE sur l'ensemble d'apprentissage

Ici nous calculerons directement l'erreur quadratique sur l'ensemble d'apprentissage sans passer par la décomposition biais variance. En utilisant l'approximation décrite en section 5.2.1 nous pouvons approximer la matrice $\mathbf{H}_\lambda$:

$$\mathbf{H}_\lambda = \mathbf{Z}\mathbf{K}_\lambda \simeq \begin{cases} \frac{1}{p+\lambda}\mathbf{Z}\mathbf{Z}^T \simeq \frac{p}{p+\lambda}\mathbf{I}_n \simeq h^2\mathbf{I}_n \text{ si } n<p \text{ et} \\ \frac{1}{n+\lambda}\mathbf{Z}\mathbf{Z}^T \text{ sinon.} \end{cases}$$

En supposant encore une fois que $\mathbf{Z}$ est fixée et en écrivant l'espérance selon $\mathbf{y}_{tr}$ de l'erreur quadratique de prédiction sur l'ensemble d'apprentissage nous avons

$$\mathbb{E}_{\mathbf{y}_{tr}}\left[\frac{1}{n}(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})^T(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})\right] = \mathbb{E}_{\mathbf{y}_{tr}}\left[\frac{1}{n}\mathbf{y}_{tr}^T(\mathbf{I}_n-\mathbf{H}_\lambda)^2\mathbf{y}_{tr}\right] \tag{C.1}$$

$$= \frac{1}{n}\left(\text{tr}((\mathbf{I}_n-\mathbf{H}_\lambda)^2\times\sigma^2\mathbf{I}_n)+(\mathbf{Z}u)^T(\mathbf{I}_n-\mathbf{H}_\lambda)^2(\mathbf{Z}u)\right). \tag{C.2}$$

Concentrons nous sur l'approximation de la quantité $(\mathbf{I}_n - \mathbf{H}_\lambda)^2$:

$$(\mathbf{I}_n - \mathbf{H}_\lambda)^2 \simeq \begin{cases} (1-h^2)^2\mathbf{I}_n \text{ si } n < p \\ \mathbf{I}_n - 2 \times \frac{1}{n+\lambda}\mathbf{Z}\mathbf{Z}^T + \left(\frac{1}{n+\lambda}\right)^2 \mathbf{Z}\mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T \simeq \mathbf{I}_n - \frac{2}{n+\lambda}\mathbf{Z}\mathbf{Z}^T + \frac{n}{(n+\lambda)^2}\mathbf{Z}\mathbf{Z}^T \text{ sinon.} \end{cases}$$

En injectant cette approximation dans (C.2), nous obtenons de manière triviale une forme très simple pour le cas $n < p$. Le cas $n > p$ demande un petit peu plus de calcul :

$$\begin{aligned} \operatorname{tr}\left(\mathbf{I}_n - \frac{2}{n+\lambda}\mathbf{Z}\mathbf{Z}^T + \frac{n}{(n+\lambda)^2}\mathbf{Z}\mathbf{Z}^T\right) &= n - \frac{2}{n+\lambda}\operatorname{tr}(\mathbf{Z}\mathbf{Z}^T) + \frac{n}{(n+\lambda)^2}\operatorname{tr}(\mathbf{Z}\mathbf{Z}^T) \\ &\simeq n - 2p\frac{n}{n+\lambda} + p\left(\frac{n}{n+\lambda}\right)^2 \end{aligned}$$

$$\begin{aligned} &(\mathbf{Z}u)^T(\mathbf{Z}u) \simeq nh^2 \\ &(\mathbf{Z}u)^T(-\frac{2}{n+\lambda}\mathbf{Z}\mathbf{Z}^T)(\mathbf{Z}u) = -\frac{2}{n+\lambda}u^T\mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T\mathbf{Z}u \simeq -2n\frac{n}{n+\lambda}u^Tu \simeq -2n\frac{n}{n+\lambda}h^2 \\ &(\mathbf{Z}u)^T(\frac{n}{(n+\lambda)^2}\mathbf{Z}\mathbf{Z}^T)(\mathbf{Z}u) = \frac{n}{(n+\lambda)^2}u^T\mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T\mathbf{Z}u \simeq n\left(\frac{n}{n+\lambda}\right)^2 h^2. \end{aligned}$$

En factorisant ces résultats selon $\frac{n}{n+\lambda}$ et $\left(\frac{n}{n+\lambda}\right)^2$, nous pouvons écrire l'erreur de prédiction sur l'ensemble d'apprentissage comme

$$\mathbb{E}_{\mathbf{y}_{tr}}\left[\frac{1}{n}(\mathbf{y}_{tr} - \hat{\mathbf{y}}_{tr})^T(\mathbf{y}_{tr} - \hat{\mathbf{y}}_{tr})\right] \simeq \begin{cases} (1-h^2)^2 \text{ si} n < p \\ 1 - 2\frac{n}{n+\lambda}\left(\frac{p}{n}(1-h^2)+h^2\right) + \left(\frac{n}{n+\lambda}\right)^2\left(\frac{p}{n}(1-h^2)+h^2\right) \text{ sinon.} \end{cases}$$

Vérifions rapidement la cohérence de l'approximation : en rappelant que $\frac{n}{n+\lambda}$ est une fonction de l'héritabilité et du ration n/p (équation (5.2)) et en particulier que $\frac{n}{n+\lambda} \underset{\frac{n}{p}\to 1}{\to} h^2$,

$$\begin{aligned} &\lim_{\frac{n}{p}\to 1+} 1 - 2\frac{n}{n+\lambda}\left(\frac{p}{n}(1-h^2)+h^2\right) + \left(\frac{n}{n+\lambda}\right)^2\left(\frac{p}{n}(1-h^2)+h^2\right) \\ &\quad = 1 - 2h^2(1-h^2+h^2) + (h^2)^2(1-h^2+h^2) \\ &\quad = 1 - 2h^2 + (h^2)^2 = (1-h^2)^2. \end{aligned}$$

C.3 Calcul du carré de la corrélation sur un individu de test

La formule pour la corrélation entre un individu de test et sa prédiction est

$$\text{corr}^2(y_{te}, \hat{y}_{te}) = \frac{\text{cov}^2(y_{te}, \hat{y}_{te})}{\text{var}\,[y_{te}]\,\text{var}\,[\hat{y}_{te}]}.$$

Nous mettrons l'aléa sur y_{te}, z_{te}, et y_{tr} et utiliserons les mêmes hypothèses et approximations que pour l'erreur de prédiction.

Écrivons chacun des trois termes de la covariance :

$$\begin{aligned}
\text{var}_{\mathbf{y}_{tr},y_{te},z_{te}}\,[\hat{y}_{te}] &= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[\hat{y}_{te}^2] - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[\hat{y}_{te}]^2 \\
&= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}\mathbf{y}_{tr}^T\mathbf{K}_\lambda^T z_{te}] - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}]^2 \\
&= \mathbb{E}_{\mathbf{y}_{tr}}\mathbb{E}_{y_{te},z_{te}|\mathbf{y}_{tr}}[z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}\mathbf{y}_{tr}^T\mathbf{K}_\lambda^T z_{te}] - \mathbb{E}_{\mathbf{y}_{tr}}\mathbb{E}_{y_{te},z_{te}|\mathbf{y}_{tr}}[z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}]^2 \\
&= \mathbb{E}_{\mathbf{y}_{tr}}[\text{tr}(\mathbf{K}_\lambda\mathbf{y}_{tr}\mathbf{y}_{tr}^T\mathbf{K}_\lambda^T) + 0] - 0 \\
&= \mathbb{E}_{\mathbf{y}_{tr}}[\mathbf{y}_{tr}^T\mathbf{K}_\lambda^T\mathbf{K}_\lambda\mathbf{y}_{tr}] \\
&= \text{tr}(\mathbf{K}_\lambda^T\mathbf{K}_\lambda \times \sigma^2\mathbf{I}_n) + (\mathbf{Z}u)^T\mathbf{K}_\lambda^T\mathbf{K}_\lambda(\mathbf{Z}u).
\end{aligned}$$

$$\begin{aligned}
\text{cov}_{\mathbf{y}_{tr},y_{te},z_{te}}(y_{te}, \hat{y}_{te}) &= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[(y_{te} - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[y_{te}])(\hat{y}_{te} - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[\hat{y}_{te}])] \\
&= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[(y_{te} - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[z_{te}^T u + e_{te}])(\hat{y}_{te} - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}])] \\
&= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[(y_{te}\hat{y}_{te})] \\
&= \mathbb{E}_{y_{te},z_{te}}\mathbb{E}_{\mathbf{y}_{tr}|y_{te},z_{te}}[y_{te}z_{te}^T\mathbf{K}_\lambda\mathbf{y}_{tr}] \\
&= \mathbb{E}_{y_{te},z_{te}}[y_{te}z_{te}^T\mathbf{K}_\lambda\mathbf{Z}u] \\
&= \mathbb{E}_{y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{Z}u(u^T z_{te} + e_{te}^T)] \\
&= \mathbb{E}_{y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{Z}uu^T z_{te}] + \mathbb{E}_{y_{te},z_{te}}[z_{te}^T\mathbf{K}_\lambda\mathbf{Z}u \times e_{te}^T] \\
&= \text{tr}(\mathbf{K}_\lambda\mathbf{Z}uu^T) + 0 + 0 \ \left(z_{te}^T \perp e_{te},\ \mathbb{E}[e_{te}] = 0\right) \\
&= u^T\mathbf{K}_\lambda\mathbf{Z}u.
\end{aligned}$$

$$\begin{aligned}
\text{var}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[y_{te}\right] &= \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[y_{te}^2] - \mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}[y_{te}]^2 \\
&= \mathbb{E}_{y_{te},z_{te}}[(z_{te}^T u + e_{te})^2] - 0 \\
&= \mathbb{E}_{e_{te},z_{te}}\left[(z_{te}^T u)^2\right] + \mathbb{E}_{e_{te},z_{te}}\left[(e_{te})^2\right] + 2\mathbb{E}_{e_{te},z_{te}}\left[(z_{te}^T u)e_{te}\right] \\
&= \mathbb{E}_{z_{te}}[z_{te}^T u u^T z_{te}] + \mathbb{E}_{e_{te}}[e_{te}^2] + 2\mathbb{E}_{z_{te}}[z_{te}^T u]\mathbb{E}_{e_{te}}[e_{te}] \\
&= u^T u + \sigma^2 + 0.
\end{aligned}$$

En résumé avons donc :

$$\begin{aligned}
\text{cov}^2(y_{te}, \hat{y}_{te}) &= (u^T \mathbf{K}_\lambda \mathbf{Z}_{tr} u)^2, \\
\text{var}\left[\hat{y}_{te}\right] &= \text{tr}(\mathbf{K}_\lambda^T \mathbf{K}_\lambda \times \sigma^2 \mathbf{I}_n) + (\mathbf{Z}_{tr} u)^T \mathbf{K}_\lambda^T \mathbf{K}_\lambda (\mathbf{Z}_{tr} u), \\
\text{var}\left[y_{te}\right] &= 1.
\end{aligned}$$

Approximons maintenant ces quantités en approchant $\mathbf{Z}\mathbf{Z}^T$ ou $\mathbf{Z}^T\mathbf{Z}$ selon la valeur du ratio n/p.

Quand $n < p$,

$$\mathbf{Z}\mathbf{Z}^T \simeq p\mathbf{I}_n \Rightarrow \mathbf{K}_\lambda \simeq \frac{1}{p+\lambda}\mathbf{Z}^T \Rightarrow \mathbf{K}_\lambda^T \mathbf{K}_\lambda \simeq \frac{(h^2)^2}{p}\mathbf{I}_n$$

et donc

$$\begin{aligned}
&\sigma^2 \times \text{tr}(\mathbf{K}_\lambda^T \mathbf{K}_\lambda) \simeq \frac{n}{p}(h^2)^2(1-h^2) \\
&(\mathbf{Z}u)^T \mathbf{K}_\lambda^T \mathbf{K}_\lambda (\mathbf{Z}u) \simeq \frac{(h^2)^2}{p}(\mathbf{Z}u)^T(\mathbf{Z}u) \simeq \frac{n}{p}(h^2)^2 \times h^2 \\
&u^T \mathbf{K}_\lambda \mathbf{Z} u \simeq u^T \frac{1}{p+\lambda}\mathbf{Z}^T\mathbf{Z}u \simeq \frac{1}{p+\lambda} n h^2 = \frac{n}{p}(h^2)^2.
\end{aligned}$$

A l'inverse quand $n > p$

$$\mathbf{Z}^T\mathbf{Z} \simeq n\mathbf{I}_p \Rightarrow \mathbf{K}_\lambda \simeq \frac{1}{n+\lambda}\mathbf{Z}^T \Rightarrow \mathbf{K}_\lambda^T \mathbf{K}_\lambda \simeq \left(\frac{1}{n+\lambda}\right)^2 \mathbf{Z}\mathbf{Z}^T$$

et nous pouvons ainsi écrire

$$\text{tr}(\mathbf{K}_\lambda^T\mathbf{K}_\lambda \times \sigma^2\mathbf{I}_n) \simeq (1-h^2)\left(\frac{1}{n+\lambda}\right)^2 \text{tr}(\mathbf{Z}\mathbf{Z}^T) \simeq (1-h^2)\frac{n}{(n+\lambda)^2}p = (1-h^2)\left(\frac{n}{n+\lambda}\right)^2\frac{p}{n}$$

$$(\mathbf{Z}u)^T\mathbf{K}_\lambda^T\mathbf{K}_\lambda(\mathbf{Z}u) \simeq \frac{1}{(n+\lambda)^2}u^T\mathbf{Z}^T\mathbf{Z}\mathbf{Z}^T\mathbf{Z}u \simeq \left(\frac{n}{n+\lambda}\right)^2 u^Tu \simeq \left(\frac{n}{n+\lambda}\right)^2 h^2$$

$$u^T\mathbf{K}_\lambda\mathbf{Z}u \simeq \frac{n}{n+\lambda}h^2.$$

En concaténant tous ces résultats nous obtenons

$$\text{corr}(\hat{y}_{te}, y_{te}) \simeq \begin{cases} \frac{\frac{n}{p}(h^2)^2}{\sqrt{\frac{n}{p}(h^2)^2\sqrt{1}}} = \sqrt{\frac{n}{p}}h^2 \text{ si } n < p \\ \frac{\frac{n}{n+\lambda}h^2}{\sqrt{\left(\frac{n}{n+\lambda}\right)^2\left(\frac{p}{n}(1-h^2)+h^2\right)\sqrt{1}}} = \frac{h^2}{\sqrt{\frac{p}{n}(1-h^2)+h^2}}, \quad \text{sinon.} \end{cases} \tag{C.3}$$

et il ne reste plus qu'à tout mettre au carré.

C.4 Calcul de l'estimateur de la variance des coefficients de $\hat{u}_R$

La formule de la variance du vecteur d'estimation des effets est

$$\begin{aligned} \text{var}(\hat{u}_R) &= \frac{1}{p}\mathbb{E}_\mathbf{y}\left[(\hat{u}_R - \mathbb{E}_\mathbf{y}[\hat{u}_R])^T(\hat{u}_R - \mathbb{E}_\mathbf{y}[\hat{u}_R])\right] \\ &= \frac{1}{p}\mathbb{E}_\mathbf{y}\left[(\mathbf{K}_\lambda\mathbf{e})^T(\mathbf{K}_\lambda\mathbf{e})\right] \\ &= \frac{1}{p}\left(\text{tr}(\sigma^2\mathbf{K}_\lambda^T\mathbf{K}_\lambda) + 0\right). \end{aligned}$$

Rappelons que $\mathbf{K}_\lambda = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I}_n)^{-1} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I}_p)^{-1}\mathbf{Z}^T$, nous avons donc

$$\mathbf{K}_\lambda = \begin{cases} \frac{1}{p+\lambda}\mathbf{Z}^T \text{ si } n < p \\ \frac{1}{n+\lambda}\mathbf{Z}^T \text{ sinon.} \end{cases}$$

Ainsi nous pouvons écrire l'approximation de la formule de variance selon le ratio n/p :

$$\text{var}(\hat{u}_R) = \begin{cases} \frac{\sigma^2}{p} \times \left(\frac{1}{p+\lambda}\right)^2 \times p \times n = \frac{1}{p}(1-h^2)(h^2)^2\frac{n}{p} \text{ si } n < p \\ \frac{\sigma^2}{p} \times \left(\frac{1}{n+\lambda}\right)^2 \times p \times n = \frac{1}{p}(1-h^2)\left(\frac{n}{n+\lambda}\right)^2\frac{p}{n} \text{ sinon.} \end{cases}$$

C.5 Pourcentage d'individus de l'ensemble d'apprentissage passant les filtres contrôle qualité

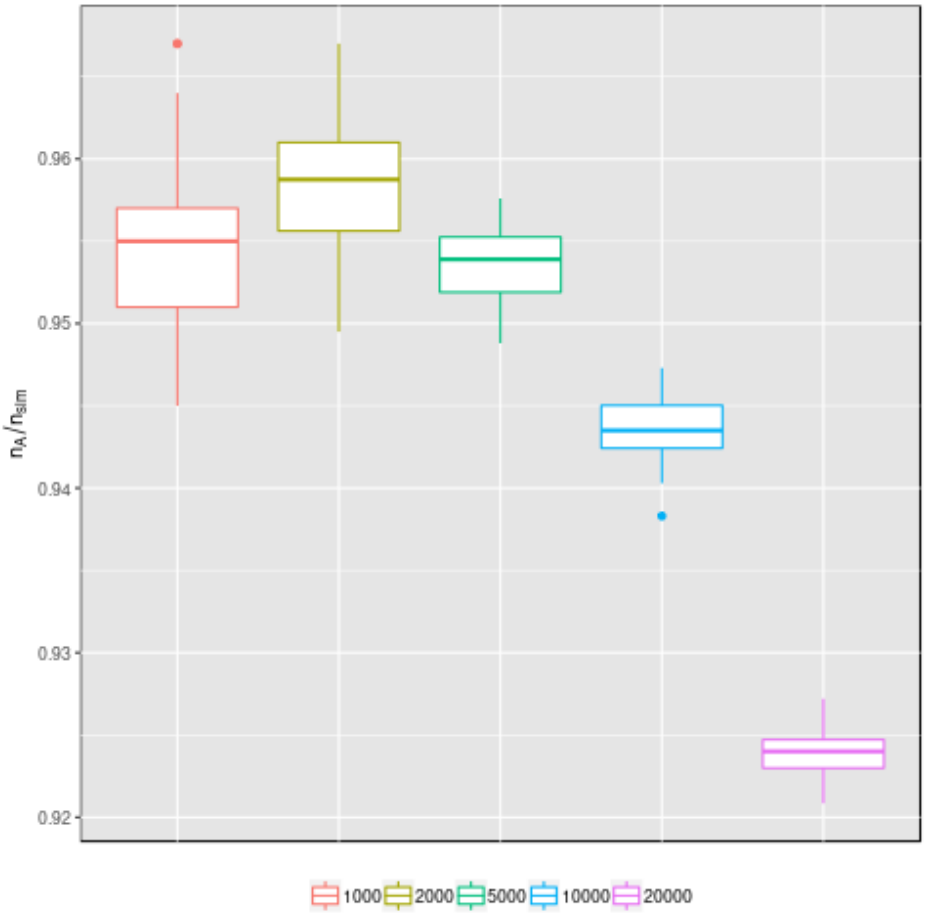


FIGURE C.1 – Graphes de proportions d'individus passant les filtres qualités. Chaque boite correspond au nombre d'individus dans nos sous-échantillon avant les différents filtres et l'ordonnée correspond au ratio nombre d'individus après filtre sur nombre d'individus avant filtre.

C.6 Estimation de pouvoir prédictifs avec paramètre de pénalisation fixé

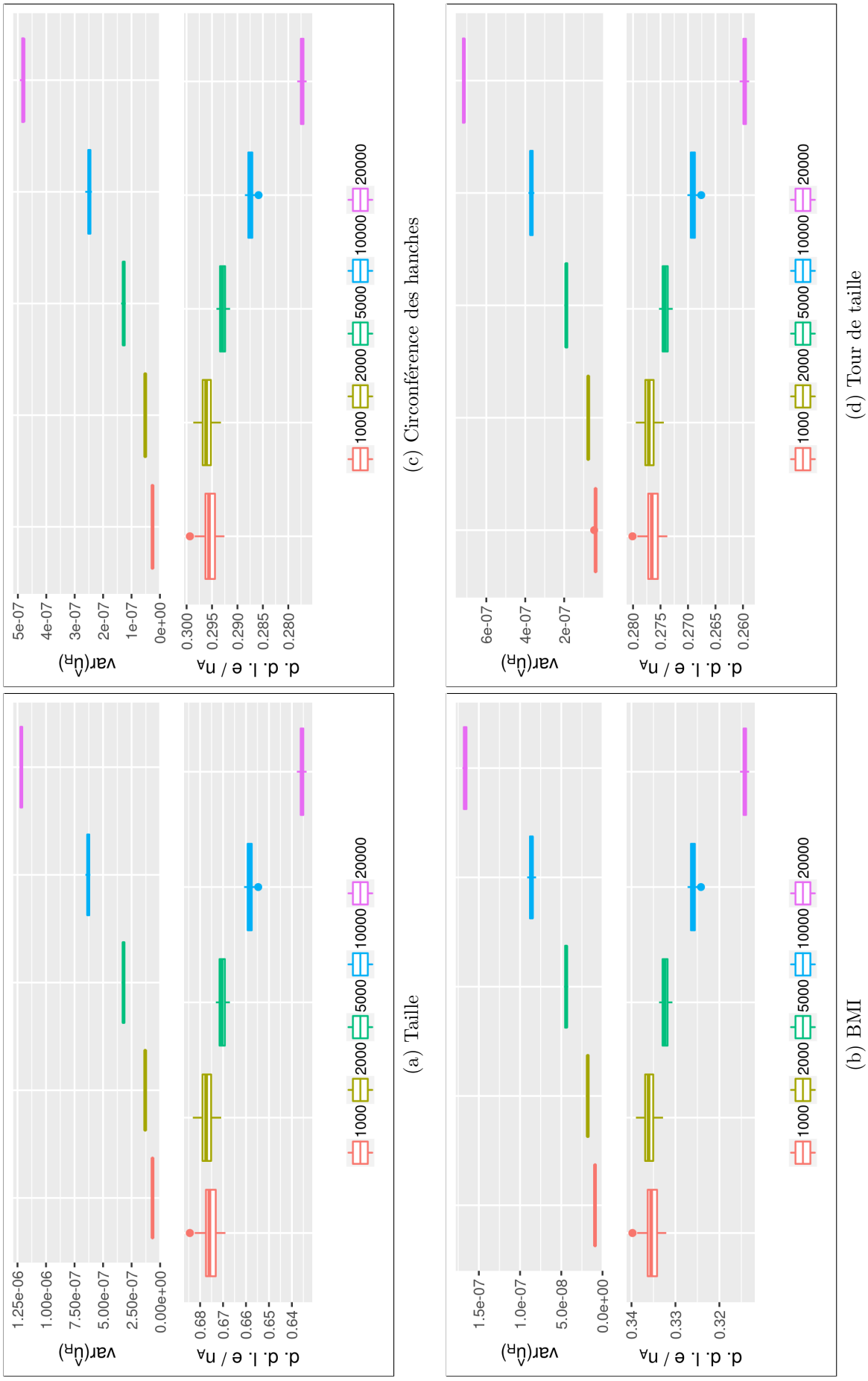
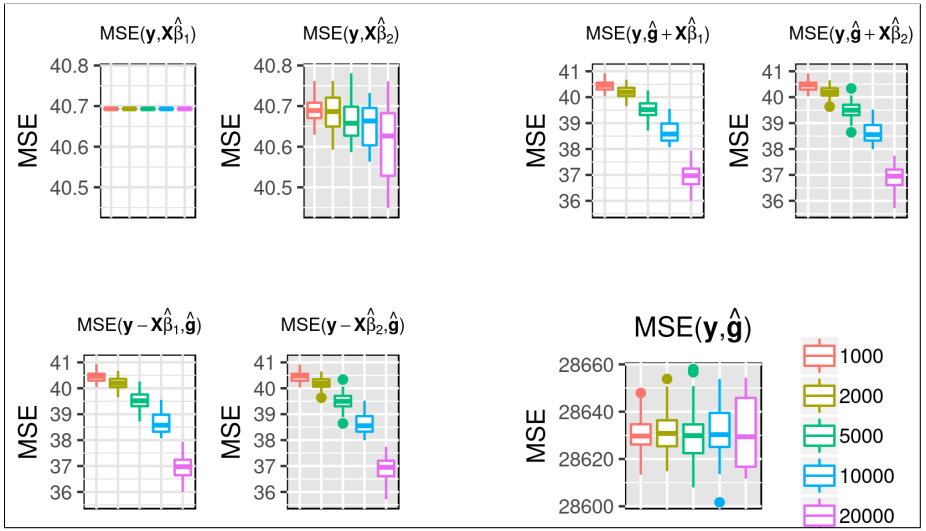
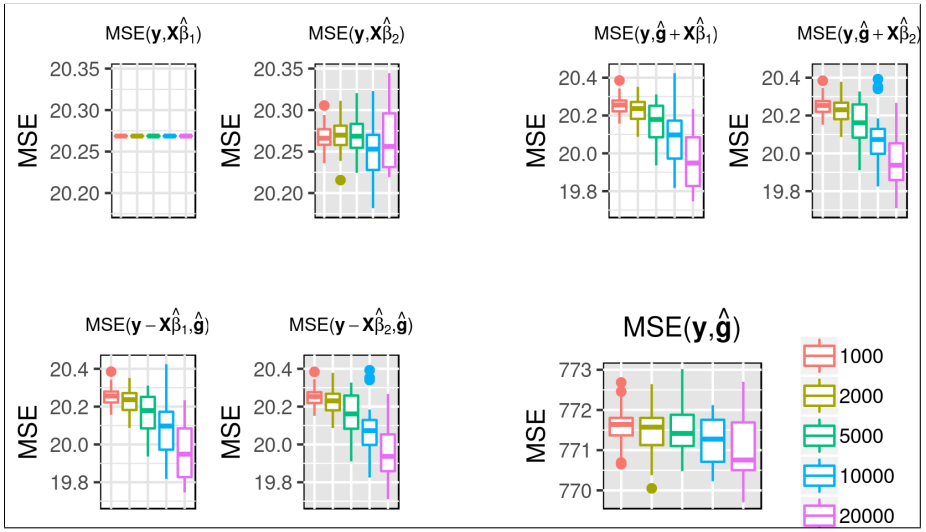


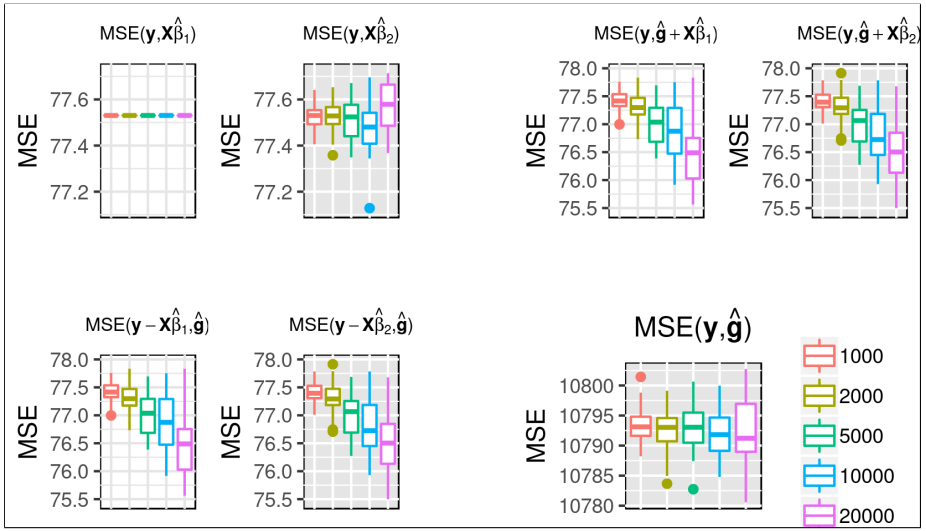
FIGURE C.2 – Graphes d'estimation d'héritabilité, de d.d.l.e et de variance empirique de $\hat{u}_R$ sur des sous-échantillons de UKBB avec héritabilité fixée. Chaque boîte correspond à un phénotype. Dans chaque boîte le graphe en haut montre les degrés de liberté effectifs et le graphe du bas la variance empirique des coefficients de $\hat{u}_R$. Les quantités sont calculées pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est fixée pour chaque phénotype à partir de valeurs empiriques. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.



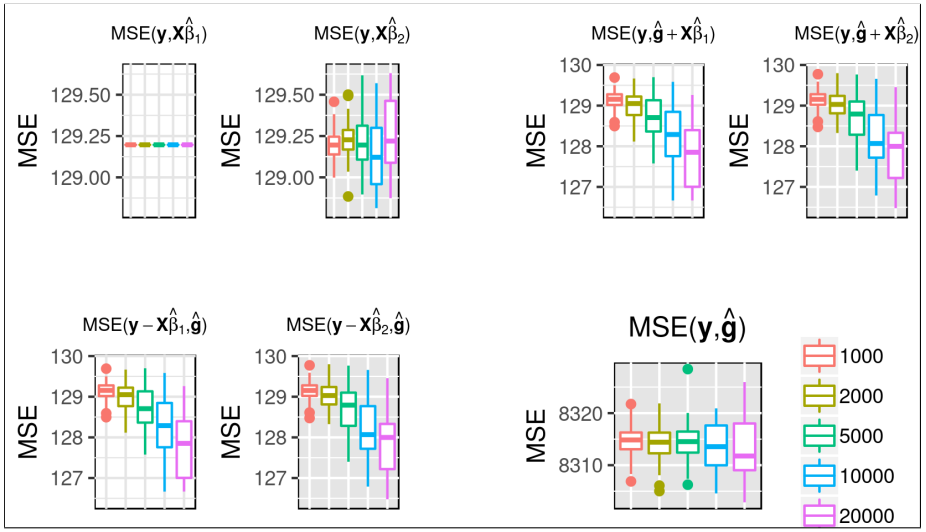
(a) Taille



(b) BMI



(c) Circonférence des hanches



(d) Tour de taille

FIGURE C.3 – Graphes des MSE estimés sur des sous-échantillons de UKBB avec héritabilité fixée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est fixée pour chaque phénotype à partir de valeurs empiriques. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.

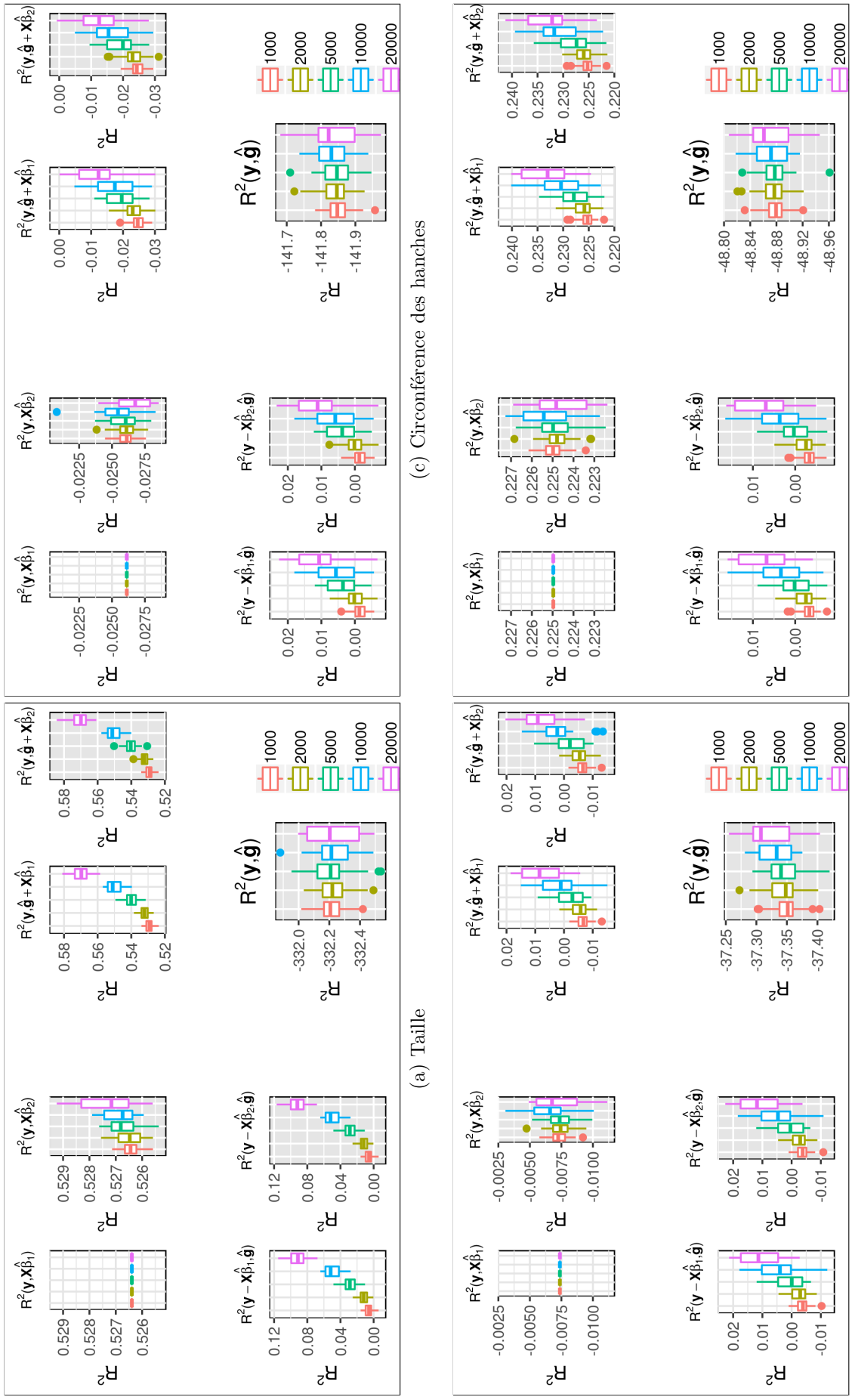
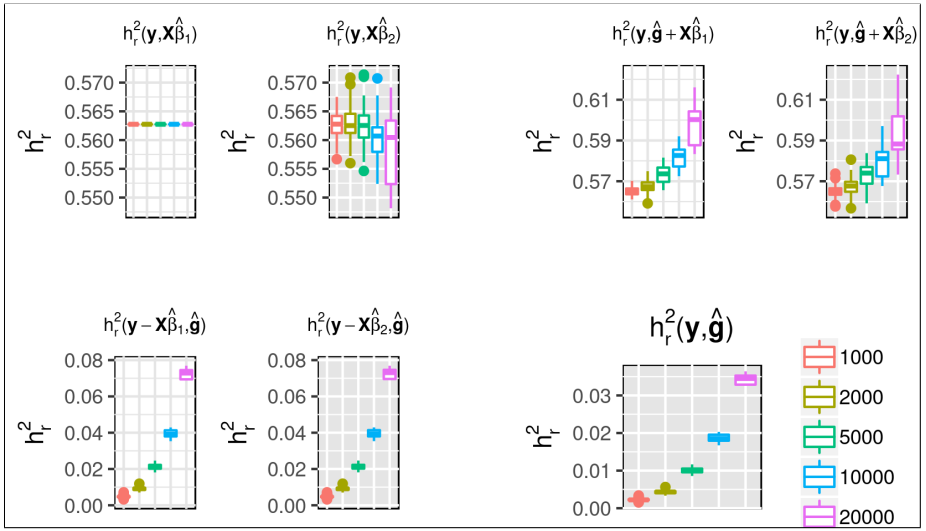
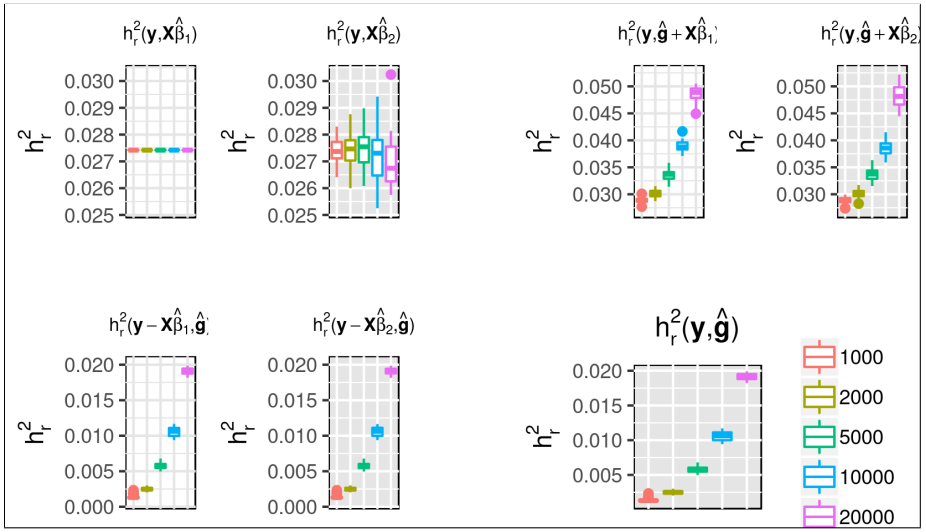


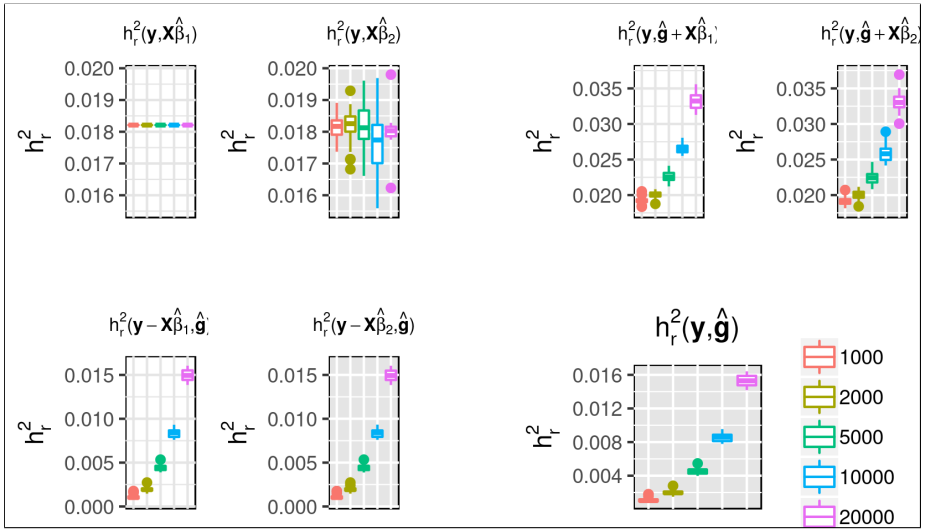
FIGURE C.4 – Graphes des R^2 estimés sur des sous-échantillons de UKBB avec héritabilité fixée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est fixée pour chaque phénotype à partir de valeurs empiriques. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.



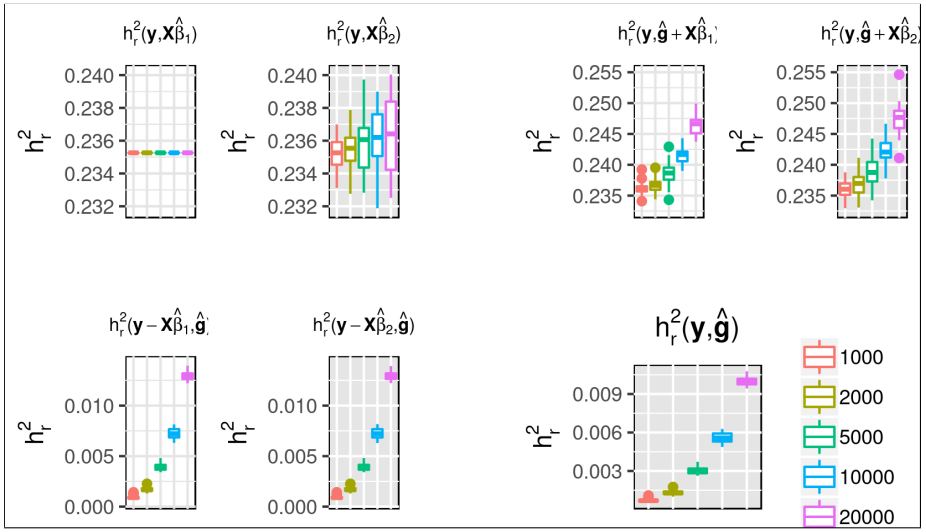
(a) Taille



(b) BMI

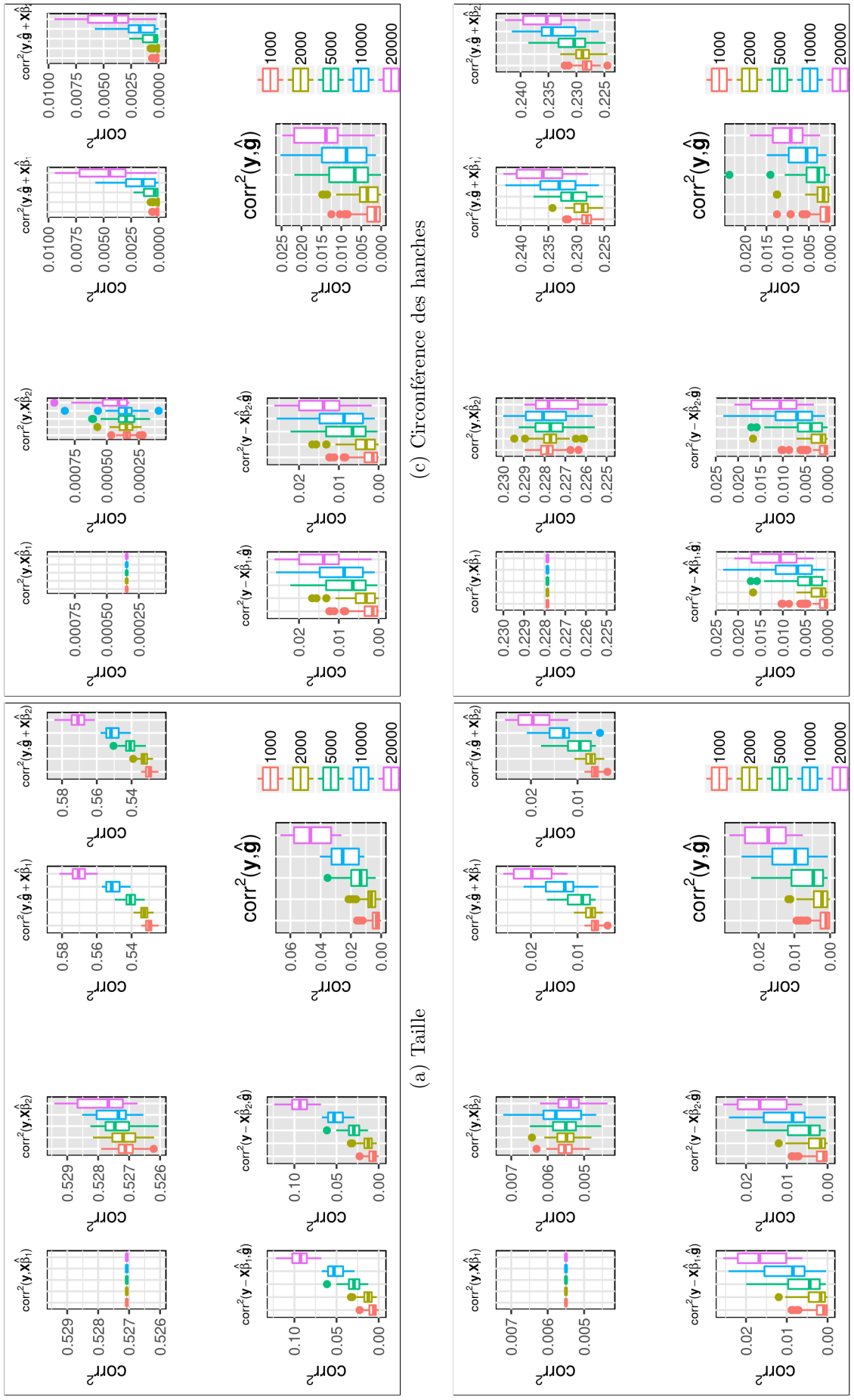


(c) Circonférence des hanches



(d) Tour de taille

FIGURE C.5 – Graphes des h_r^2 estimés sur des sous-échantillons de UKBB avec héritabilité fixée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est fixée pour chaque phénotype à partir de valeurs empiriques. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.



(a) Taille (b) BMI (c) Circonférence des hanches (d) Tour de taille

FIGURE C.6 – Graphes des corrélations estimés sur des sous-échantillons de UKBB avec héritabilité fixée. Chaque boîte correspond à un phénotype. Dans chaque boîte la quantité est calculée pour un ensemble de test de taille $n \in \{1000, 2000, 5000, 10000, 20000\}$ avec respectivement $\{1000, 70, 40, 20, 10\}$ répétitions. Pour chaque ensemble d'apprentissage l'héritabilité est fixée pour chaque phénotype à partir de valeurs empiriques. L'échantillon de test est composé de 1000 individus et celui de standardisation pour estimer les effets fixes de 1000 individus. Dans les figures $\mathbf{y}$ correspond au phénotype, $\hat{\mathbf{g}}$ correspond à l'estimation du terme génétique et $\hat{\beta}_1$, $\hat{\beta}_2$ à l'estimateur des effets fixes.

Annexe D

Estimation d'héritabilité pour le cas qualitatif

D.1 Démonstration de PCGC

D.1.1 Calcul de probabilités pour la modélisation cas-contrôle

Rappelons que nous avons posé

$$W_{ij} = \frac{(y_i - K_e)(y_j - K_e)}{K_e(1 - K_e)} \tag{D.1}$$

le produit de phénotypes normalisés et $\mathcal{S}_i$ une variable d'appartenance à l'étude : $\mathcal{S}_i = 1$ si individu i est compris dans l'étude et 0 sinon (on étend cette notation à $\mathcal{S}_{i,j} = 1$ pour l'appartenance d'une paire d'individus i et j).

Nous introduisons les probabilité suivante :

— $p_{case} = \mathbb{P}(\mathcal{S}_i = 1 \mid y_i = 1)$ la probabilité pour un cas d'être sélectionné dans l'étude.

— $p_{control} = \mathbb{P}(\mathcal{S}_i = 1 \mid y_i = 0)$ la probabilité pour un contrôle d'être sélectionné dans l'étude.

En partant de la définition mathématique de la prévalence dans l'étude et à l'aide

de la formule de Bayes, nous avons

$$K_e = \mathbb{P}(y_i = 1 \mid \mathcal{S}_i = 1) = \frac{\mathbb{P}(\mathcal{S}_i = 1 \mid y_i = 1)\mathbb{P}(y_i = 1)}{\mathbb{P}(\mathcal{S}_i = 1)} = \frac{p_{case}K}{\mathbb{P}(\mathcal{S}_i = 1)}.$$

De même nous pouvons écrire

$$1 - K_e = \frac{(1-K)p_{control}}{\mathbb{P}(\mathcal{S}_i = 1)}.$$

En combinant ces deux expressions,

$$\frac{1-K_e}{K_e} = \frac{(1-K)p_{control}}{\mathbb{P}(\mathcal{S}_i = 1)} \times \frac{\mathbb{P}(\mathcal{S}_i = 1)}{p_{case}K} \iff p_{control} = \frac{(1-K_e)K}{(1-K)K_e}p_{case}.$$

En faisant l'hypothèse que $p_{case} = 1$ i.e tout les cas que nous trouvons sont intégrés dans l'étude ("full ascertainment"), alors nous obtenons $p_{control} = \frac{(1-K_e)K}{(1-K)K_e}$. Nous en déduisons également la probabilité d'être tiré dans l'étude $\mathbb{P}(\mathcal{S}_i = 1) = \frac{K}{K_e}$.

D.1.2 L'idée de PCGC

Nous remarquons que les liability d'une paire d'individus (i, j) suivent $(l_i, l_j) \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma = \begin{pmatrix} 1 & h_l^2\mathbf{G}_{ij}^* \\ h_l^2\mathbf{G}_{ij}^* & 1 \end{pmatrix}$.

L'hypothèse de l'article de Golan et al. est que l'héritabilité contrôle le lien le phénotype et le génotype i.e.

$$\mathbb{E}(y_i y_j) = f(\mathbf{G}_{ij}^*, h_l^2)$$

avec f une fonction à déterminer. A partir de cette hypothèse, nous pouvons proposer une approximation de h_l^2,

$$\hat{h_l^2} = \underset{h^2 \in [0,1]}{\arg\min} \sum_{i<j} \left(W_{ij} - \mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}_{ij}^*)\right)^2. \tag{D.2}$$

D.1.3 Approximation de $\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}_{ij}^*)$

Nous commençons par remarquer que $W_{ij} \in \{\frac{1-K_e}{K_e}; -1; \frac{K_e}{1-K_e}\}$ selon les valeurs de y_i et y_j ((1,1),(0,1)/(1,0),(0,0)).

En utilisant la formule de l'espérance, nous pouvons alors écrire :

$$\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{1-K_e}{K_e}\mathbb{P}(y_i=y_j=1 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) - \mathbb{P}(y_i \neq y_j \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) + \frac{K_e}{1-K_e}\mathbb{P}(y_i=y_j=0 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}).$$

Exprimons ces différentes probabilité à l'aide de la formule de Bayes :

1.

$$\mathbb{P}(y_i=y_j=1 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{\mathbb{P}(y_i=y_j=1 \mid \mathbf{G}^*_{ij})\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i=y_j=1,\mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

Or $\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i=y_j=1,\mathbf{G}^*_{ij}) = {p_{case}}^2 = 1$ (full ascertainment).

$$\Rightarrow \mathbb{P}(y_i=y_j=1 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{\mathbb{P}(y_i=y_j=1 \mid \mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij} \mid \mathbf{G}^*_{ij})}.$$

2.

$$\mathbb{P}(y_i=y_j=0 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{\mathbb{P}(y_i=y_j=0 \mid \mathbf{G}^*_{ij})\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i=y_j=0,\mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

Or $\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i=y_j=0,\mathbf{G}^*_{ij}) = {p_{control}}^2 = \left[\frac{(1-K_e)K}{(1-K)K_e}\right]^2$

$$\Rightarrow \mathbb{P}(y_i=y_j=0 \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \left[\frac{(1-K_e)K}{(1-K)K_e}\right]^2 \frac{\mathbb{P}(y_i=y_j=0 \mid \mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

3.

$$\mathbb{P}(y_i \neq y_j \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{\mathbb{P}(y_i \neq y_j \mid \mathbf{G}^*_{ij})\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i \neq y_j,\mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

Or $\mathbb{P}(\mathcal{S}_{ij}=1 \mid \ y_i \neq y_j,\mathbf{G}^*_{ij}) = p_{control}p_{case} = \frac{(1-K_e)K}{(1-K)K_e}$

$$\Rightarrow \mathbb{P}(y_i \neq y_j \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{(1-K_e)K}{(1-K)K_e}\frac{\mathbb{P}(y_i \neq y_j \mid \mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

En combinant toutes ces expressions, nous pouvons alors écrire :

$$\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij}=1;\mathbf{G}^*_{ij}) = \frac{\frac{1-K_e}{K_e}\mathbb{P}(y_i=y_j=1 \mid \mathbf{G}^*_{ij}) - \frac{(1-K_e)K}{(1-K)K_e}\mathbb{P}(y_i \neq y_j \mid \mathbf{G}^*_{ij}) + \frac{K_e}{1-K_e}\left[\frac{(1-K_e)K}{(1-K)K_e}\right]^2\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}{\mathbb{P}(\mathcal{S}_{ij}=1 \mid \mathbf{G}^*_{ij})}.$$

L'idée va être d'effectuer un développement de Taylor autour de faibles valeurs de G_{ij} et de réécrire $\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}^*_{ij})$ sous forme de fraction .

$$\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}^*_{ij}) = \frac{A(\mathbf{G}^*_{ij})}{B(\mathbf{G}^*_{ij})}$$

et nous pouvons alors écrire le développement autour de $\mathbf{G}^*_{ij} = 0$:

$$\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}^*_{ij}) = \frac{A(0)}{B(0)} + \frac{A'(0)B(0) - A(0)B'(0)}{B(0)^2}\mathbf{G}^*_{ij} + \mathcal{O}\left((\mathbf{G}^*_{ij})^2\right).$$

En remarquant que $\mathbf{G}^*_{ij} = 0 \Rightarrow y_i, y_j$ i.i.d ($\Sigma = \mathbf{I}_2(\mathbb{R})$ et puisque nous sommes dans le cas gaussien $l_i \perp l_j$ et donc $y_i \perp y_j$) , alors :

— $\mathbb{E}(W_{ij}) = 0 \Rightarrow A(0) = 0$

— $\mathbb{P}(\mathcal{S}_{ij}) = \left(\frac{K}{K_e}\right)^2 \Rightarrow B(0) = \left(\frac{K}{K_e}\right)^2$.

En réécrivant le développement,

$$\mathbb{E}(W_{ij} \mid \mathcal{S}_{ij} = 1; \mathbf{G}^*_{ij}) = \left(\frac{K_e}{K}\right)^2 A'(0)\mathbf{G}^*_{ij} + \mathcal{O}\left((\mathbf{G}^*_{ij})^2\right).$$

La paire de liability $(l_i, l_j) \sim \mathcal{N}(0, \Sigma)$ peut être vu comme la somme de deux gaussiennes. A partir de la définition de la liability ($l_i > t \Leftrightarrow y_i = 1$), on peut donc écrire

$$\mathbb{P}(y_i = \mathbb{1}_{l_i > t}, y_j = \mathbb{1}_{l_j > t} \mid \mathbf{G}^*_{ij}) = \int_{L_i}\int_{L_j} f_\Sigma(l_1, l_2) dl_1 dl_2 \tag{D.3}$$

avec :

— $f_\Sigma(l_1, l_2) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{(l_1,l_2)\Sigma^{-1}(l_1,l_2)^T}{2}}$ la densité d'une gaussienne à deux dimension et à moyenne nulle.

— $\forall i, l_i =]-\infty, t[$ si $y_i = 0$ et $[t, +\infty[$ sinon.

Nous avons $|\Sigma| = 1 - (h^2_l \mathbf{G}^*_{ij})^2$, $\Sigma^{-1} = \begin{pmatrix} 1 & -h^2_l \mathbf{G}^*_{ij} \\ -h^2_l \mathbf{G}^*_{ij} & 1 \end{pmatrix}$ et nous pouvons donc écrire

$$f_\Sigma(l_1, l_2) = \frac{1}{2\pi\sqrt{1 - (h^2_l \mathbf{G}^*_{ij})^2}} e^{-\frac{l_1^2 + l_2^2 - 2l_1 l_2 h^2_l \mathbf{G}^*_{ij}}{2\left[1 - (h^2_l \mathbf{G}^*_{ij})^2\right]}}.$$

Nous allons chercher une expression de $A'(\mathbf{G}^*_{ij})$ et pour cela nous devons dériver chacun des termes à double intégrales.

$$\frac{d}{d_{\mathbf{G}^*_{ij}}} \int_{L_i}\int_{L_j} f_\Sigma(l_1,l_2)dl_1dl_2 = \int_{L_i}\int_{L_j} \frac{d}{d_{\mathbf{G}^*_{ij}}} f_\Sigma(l_1,l_2)dl_1dl_2$$

$$\frac{d}{d_{\mathbf{G}^*_{ij}}} f_\Sigma(l_1,l_2) = \frac{\mathbf{G}^*_{ij}(h^2_l)^2}{2\pi(1-(\mathbf{G}^*_{ij}h^2_l)^2)^{3/2}} e^{-\frac{l_1^2+l_2^2-2l_1l_2h^2_l\mathbf{G}^*_{ij}}{2\left[1-(h^2_l\mathbf{G}^*_{ij})^2\right]}} + \frac{1}{2\pi\sqrt{1-(\mathbf{G}^*_{ij}h^2_l)^2}} \frac{2l_1l_2h^2_l\left[1-(\mathbf{G}^*_{ij}h^2_l)^2\right]+2\mathbf{G}^*_{ij}(h^2_l)^2\left[l_1^2+l_2^2-2l_1l_2h^2_l\mathbf{G}^*_{ij}\right]}{\left[1-(h^2_l\mathbf{G}^*_{ij})^2\right]^2} e^{-\frac{l_1^2+l_2^2-2l_1l_2h^2_l\mathbf{G}^*_{ij}}{2\left[1-(h^2_l\mathbf{G}^*_{ij})^2\right]}}$$

Après réécriture nous avons

$$\frac{d}{d_{\mathbf{G}^*_{ij}=0}} f_\Sigma(l_1,l_2) = \frac{1}{2\pi} l_1l_2h^2_l e^{-\frac{l_1^2+l_2^2}{2}}$$

et donc

$$\int_{L_i}\int_{L_j} \frac{d}{d_{\mathbf{G}^*_{ij}}} f_\Sigma(l_1,l_2)dl_1dl_2 = h^2_l\left[\int_{L_i}\frac{l_1}{\sqrt{2\pi}}e^{-l_1^2/2}dl_1\right]\left[\int_{L_j}\frac{l_2}{\sqrt{2\pi}}e^{-l_2^2/2}dl_2\right].$$

Or $\frac{l}{\sqrt{2\pi}}e^{-l^2/2} = -\phi'(l)$ avec ϕ la densité de la loi normale : nous pouvons alors écrire

— $\frac{d}{d_{\mathbf{G}^*_{ij}=0}}\mathbb{P}(y_i = y_j = 1 \mid \mathcal{S}_{ij}=1,\mathbf{G}^*_{ij}) = h^2_l\left[-\phi(+\infty)+\phi(t)\right]\left[-\phi(+\infty)+\phi(t)\right] = h^2_l\phi(t)$

— $\frac{d}{d_{\mathbf{G}^*_{ij}=0}}\mathbb{P}(y_i = y_j = 0 \mid \mathcal{S}_{ij}=1,\mathbf{G}^*_{ij}) = h^2_l\left[-\phi(t)+\phi(+\infty)\right]\left[-\phi(t)+\phi(+\infty)\right] = h^2_l\phi(t)$

— $\frac{d}{d_{\mathbf{G}^*_{ij}=0}}\mathbb{P}(y_i \neq y_j \mid \mathcal{S}_{ij}=1,\mathbf{G}^*_{ij}) = h^2_l\left[-\phi(+\infty)+\phi(t)\right]\left[-\phi(t)+\phi(+\infty)\right] = -h^2_l\phi(t).$

En intégrant ces expressions dans $A'(0)$ nous obtenons

$$\begin{aligned}A'(0) &= \frac{d}{d_{\mathbf{G}^*_{ij}=0}}\left[\frac{K_e}{1-K_e}\mathbb{P}(y_i=y_j=1\mid\mathcal{S}_{ij}=1,\mathbf{G}^*_{ij}) - \frac{(1-K_e)K}{(1-K)K_e}\mathbb{P}(y_i\neq y_j\mid\mathcal{S}_{ij}=1,\mathbf{G}^*_{ij}) + \frac{1-K_e}{K_e}\left(\frac{(1-K_e)K}{(1-K)K_e}\right)^2\mathbb{P}(y_i=y_j=0\mid\mathcal{S}_{ij}=1,\mathbf{G}^*_{ij})\right]\\ &= \left[\frac{K_e}{1-K_e}+2\frac{(1-K_e)K}{(1-K)K_e}+\frac{1-K_e}{K_e}\left(\frac{(1-K_e)K}{(1-K)K_e}\right)^2\right]h^2_l\phi(t)^2\\ &= \frac{1-K_e}{(1-K)^2K_e}h^2_l\phi(t)^2.\end{aligned}$$

En combinant tous ces résultats, autour des faibles valeurs de $\mathbf{G}^*_{ij}$ nous avons l'approximation :

$$\begin{aligned}\mathbb{E}\left(W_{ij} \mid \mathcal{S}_{ij} = 1, \mathbf{G}^*_{ij}\right) &\simeq \frac{A'(0)}{B(0)}\mathbf{G}^*_{ij} \\ &= \frac{K_e(1-K_e)}{K^2(1-K)^2} h_l^2 \phi(t)^2 \mathbf{G}^*_{ij} \\ &= c h_l^2 \mathbf{G}^*_{ij}\end{aligned}$$

D.2 Estimation des variances des composantes du modèle

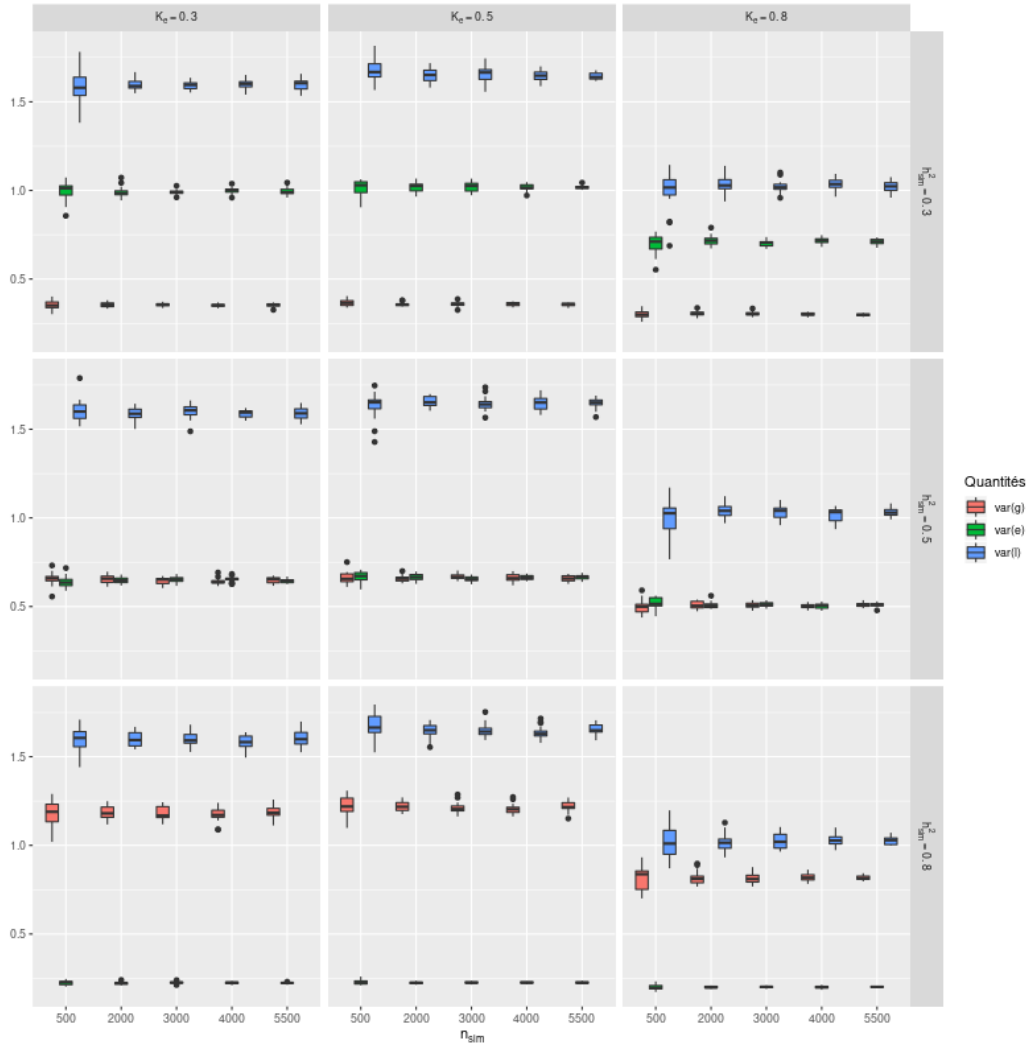


FIGURE D.1 – Graphe d'influence de n, h^2_{sim} et K_e sur l'estimation des composantes de variance. Les lignes et colonnes de la grille de boxplots correspondent respectivement aux différentes valeurs d'h^2_{sim} et de K_e. Dans chacun des boxplots, l'axe des abscisses correspond aux différentes valeurs de n et l'axes des ordonnées aux variances des termes génétique (rouge), environnemental (vert) et de liability (bleu).

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre: Lien entre héritabilité et prédiction de phénotypes complexes chez l'humain : une approche du problème par la régression ridge sur des données de population

Mots clés: Apprentissage statistique, Grande dimension, Régression ridge, Héritabilité des phénotypes humains complexes, Prédiction de phénotypes complexes, Données de génotypage à haut débit

Résumé: Cette thèse étudie l'apport des méthodes d'apprentissage automatique pour la prédiction de phénotypes humains complexes et héritables, à partir de données génétiques en population. En effet, les études d'association à l'échelle du génome (GWAS) n'expliquent en général qu'une petite fraction de l'héritabilité observée sur des données familiales. Cependant l'héritabilité peut être approchée sur des données de population par l'héritabilité génomique, qui estime la variance phénotypique expliquée par l'ensemble des polymorphismes nucléotidiques (SNP) du génome à l'aide de modèles mixtes. Cette thèse aborde donc l'héritabilité du point de vue de l'apprentissage automatique et examine le lien étroit entre les modèles mixtes et la régression ridge. Notre contribution est double. Premièrement, nous proposons d'estimer l'héritabilité génomique en utilisant une approche prédictive via la régression ridge et la validation croisée généralisée (GCV). Deuxièmement, nous dérivons des formules simples qui expriment la précision de la prédiction par la régression ridge en fonction du rapport de la taille de la population et du nombre total de SNP, montrant clairement qu'une héritabilité élevée n'implique pas nécéssairement une prédiction précise. L'estimation de l'héritabilité via GCV et les formules de précision de prédiction sont validées à l'aide de données simulées et de données réelles de UK Biobank. La dernière partie de la thèse présente des résultats sur des phénotypes qualitatifs. Ces résultats permettent une meilleure compréhension des biais des méthodes d'estimation d'héritabilité.

Title: Link between heritability and prediction of complex phenotypes in human genomics : approaching the problem using ridge regression on population data

Keywords: Machine learning, High dimensional dataset, Ridge regression, Heritability of complex human phenotypes , Prediction of complex phenotypes, High-throughput genotyping data

Abstract: This thesis studies the contribution of machine learning methods for the prediction of complex and heritable human phenotypes, from population genetic data. Indeed, genome-wide association studies (GWAS) generally only explain a small fraction of the heritability observed in family data. However, heritability can be approximated on population data by genomic heritability, which estimates the phenotypic variance explained by the set of single nucleotide polymorphisms (SNPs) of the genome using mixed models. This thesis therefore approaches heritability from a machine learning perspective and examines the close link between mixed models and ridge regression. Our contribution is twofold. First, we propose to estimate genomic heritability using a predictive approach via ridge regression and generalized cross validation (GCV). Second, we derive simple formulas that express the precision of the ridge regression prediction as a function of the size of the population and the total number of SNPs, showing that a high heritability does not necessarily imply an accurate prediction. Heritability estimation via GCV and prediction precision formulas are validated using simulated data and real data from UK Biobank. The last part of the thesis presents results on qualitative phenotypes. These results allow a better understanding of the biases of the heritability estimation methods.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France